CCF Transactions on Pervasive Computing and Interaction (2022) 4:219–239
https://doi.org/10.1007/s42486-022-00108-3

REGULAR PAPER

AviPer: assisting visually impaired people to perceive the world with visual-tactile multimodal attention network

Xinrong Li[1,2] · Meiyu Huang[1] · Yao Xu[1] · Yingze Cao[1] · Yamei Lu[1] · Pengfei Wang[1] · Xueshuang Xiang[1]

Received: 2 January 2022 / Accepted: 27 March 2022 / Published online: 30 June 2022

Abstract

Unlike able-bodied persons, it is difficult for visually impaired people, especially those in the educational age, to build a full perception of the world due to the lack of normal vision. The rapid development of AI and sensing technologies has provided new solutions to visually impaired assistance. However, to our knowledge, most previous studies focused on obstacle avoidance and environmental perception but paid less attention to educational assistance for visually impaired people. In this paper, we propose AviPer, a system that aims to assist visually impaired people to perceive the world via creating a continuous, immersive, and educational assisting pattern. Equipped with a self-developed flexible tactile glove and a webcam, AviPer can simultaneously predict the grasping object and provide voice feedback using the vision-tactile fusion classification model, when a visually impaired people is perceiving the object with his gloved hand. To achieve accurate multimodal classification, we creatively embed three attention mechanisms, namely temporal, channel-wise, and spatial attention in the model. Experimental results show that AviPer can achieve an accuracy of 99.75% in classification of 10 daily objects. We evaluated the system in a variety of extreme cases, which verified its robustness and demonstrated the necessity of visual and tactile modal fusion. We also conducted tests in the actual use scene and proved the usability and user-friendliness of the system. We opensourced the code and self-collected datasets in the hope of promoting research development and bringing changes to the lives of visually impaired people.

Keywords Assistance for visually impaired people · Multimodal learning · Visual sense · Tactile sense · Deep learning · Attention mechanism

1 Introduction

Visual impairment refers to the loss of visual acuity that cannot be ameliorated by refractive correction or medical technologies (Rahman et al. 2021). It is difficult to build a full perception of the world for visually impaired people especially those at the phase of education, due to the lack of a sane vision, which ultimately affects their living abilities, self-esteem, and mental health. According to Organization et al. (2019), as of 2010, there are 285 million visually impaired people in the world, among which 37 million are blind. And based on the estimation of Ackland et al. (2017), the number of blind will reach 55 million by 2030, and 115 million by 2050. What is sad is that, except for a few visual impairments caused by senile eye diseases such as cataracts and glaucoma, the vast majority of visual impairments are congenital, which means that these people have never seen the world they live in their whole lives. Due to medical limitations, there is no easy cure for most visual

Xinrong Li and Meiyu Huang contributed equally to this work.

✉ Xueshuang Xiang
xiangxueshuang@qxslab.cn

Xinrong Li
xr-li19@mails.tsinghua.edu.cn

Meiyu Huang
huangmeiyu@qxslab.cn

Yao Xu
xuyao@qxslab.cn

Yingze Cao
caoyingze@qxslab.cn

Yamei Lu
ymll0120@163.com

Pengfei Wang
wangpengfei@qxslab.cn

[1] Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China

[2] School of Aerospace, Tsinghua University, Beijing, China

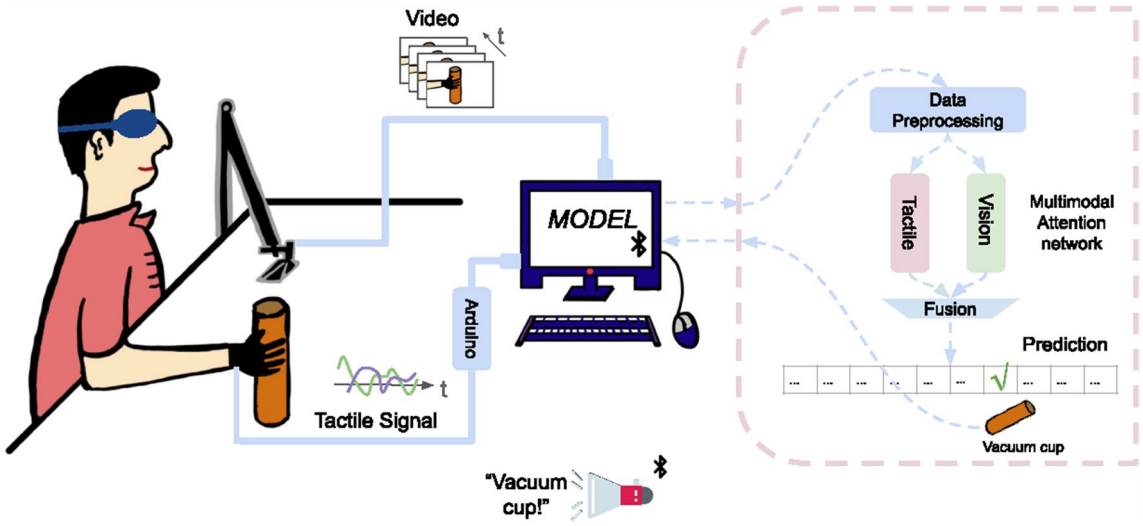


Fig. 1 System overview. The user sits at the table and grasps the object to be identified with his hand wearing the tactile glove. The glove collects the tactile time series and transmits the data to the computer through the USB interface of two Arduino Mega microcontrollers. At the same time, the webcam arranged on the desktop transmits the video of the hand area to the computer. After data preprocessing, the tactile data and visual data are sent into the trained multi-modal attention-based classification model. Finally, the model gives the classification label and broadcasts it to the user through a Bluetooth speaker

impairments. Hence, how to help visually impaired people integrate into society has become an urgent problem to be solved. Although visually impaired people who are educated in special education schools can obtain basic living abilities, not all visually impaired people have such opportunities and conditions. There has been a lack of sufficient professional educators for special education positions. In addition, most special education schools in poor areas lack the budget to hire professional special education workers, which has led to not only the lack of assistance for visually impaired people, but also more potential discrimination, abuse, and indecency against them (Warren 1994).

With the rapid development of artificial intelligence and advanced sensing technology, the use of deep learning to assist visually impaired people is booming. In recent years, there has been lots of work using deep learning to help visually impaired people perceive the environment and avoid obstacles (Poggi and Mattoccia 2016; Liu et al. 2021; Kumar et al. 2019). However, to our knowledge, research on deep learning based educational assistance systems targeted at visually impaired people is insufficient. In the field of human–computer interaction, paper Metatla et al. (2020) used a co-design approach to design and evaluate a robot-based educational game that could be inclusive of both visually impaired and sighted children. But it cannot continuously educate visually impaired people to develop their ability to perceive the world. Paper Ahmetovic et al. (2020) provides a vision-based deep learning approach to assist visually impaired people to recognize daily objects. However, due to that the classification algorithm is just vision-based, visually impaired people are difficult to get perceptual feedback and a sense of participation in the whole process. In order to bridge the gap in this area, we proposed AviPer, a system that aims to assist visually impaired people to perceive the world with visual and tactile multimodal object classification.

Our objective is to develop a continuous, immersive, and educational assistance for people who are visually impaired. More specifically, we expect to create a pattern that visually impaired people can safely and continuously learn to identify living objects without being supervised or taught, while they can have as much experience as possible about the objects they are learning to recognize. These demands lead us to multi-modal deep learning. Multimodal learning, or multi-view learning, refers to building models that can process and relate information from multiple modalities (Baltrušaitis et al. 2018). We integrate the recognition of tactile signals with visual recognition, in order to attach more classification evidence for the model, and provide users with a real sense of grasping operation. In the evaluation experiment, we proved that the visual and tactile modal fusion is very necessary and beneficial, as the multimodal model can achieve robust classification in extreme cases which are hard to distinguish with only one modality. Besides, we innovatively embedded three kinds of different attention mechanisms, namely temporal, spatial, and channel-wise attention, to better extract important information from the whole process of grabbing objects for classification. Figure 1 shows the whole process of the system.

To unleash the power of visual and tactile multimodal attention network in the application scenario of assisting visually impaired people, we have to address a series of challenges:

1. *Accurate tactile data acquisition* In previous studies, deep learning tasks based on tactile sensors are often deployed on robotic arms (Romano et al. 2011; Yuan et al. 2015; Morrison et al. 2018; Li et al. 2019). These sensors are difficult to meet the needs of wearable devices for softness and flexibility. Paper Sundaram et al. (2019) develops an advanced sensor integration, which can be fixed on a glove for wearing. But wearable device with high sensor density is not suitable for visual impairment assistance. Not only the sensor itself will incur high production costs, but also the complicated circuit requires lots of maintenance costs.
2. *Heterogeneous data* Tactile and visual signals naturally have a huge difference in scale. More specifically, vision can perceive the full view of items, while touch only reflects the local characteristics of the contact point. In a machine learning model, the gap in scale is embodied in the completely different data dimensions of the tactile signal and the video frames. Furthermore, different approaches are required to acquire the important information contained in tactile time series and video frames. Therefore, the model structure requires careful design to extract features of heterogeneous data.
3. *Privacy security* Due to the need for a webcam to obtain video frame data and the particularity of the people served, data privacy and other security issues also need to be considered.

To tactile the above challenges, we processed a tactile glove with a much lower sensor density and used our self-developed capacitance-based force sensor to *collect accurate tactile data*. The flexible sensor can directly reflect the degree of bending of the joint to help the model infer the gesture of the hand. To *handle heterogeneous data*, we creatively designed a dual-modal attention network. The model uses different modules to process tactile and visual input and conduct modality fusion with extracted feature vectors. In the modules which respectively process tactile and visual input, we implemented three kinds of attention mechanism, namely temporal, spatial and channel-wise attention to focus on key features of different modalities. In order to protect the user's *privacy security*, we strictly collect the video data of grasping action in the hand area, which means that the user will not be exposed to the webcam except for his hand with the tactile glove. In addition, the collection of data for training and testing the model is completely done by people with unimpaired vision. The rights of visually impaired people are fully respected. Hardware including tactile gloves and cameras, visual-tactile bimodal dataset, multimodal deep learning classification algorithm together constitute our assistance system for visually impaired people: AviPer. Detailed information about the system will be discussed later in Sect. 3.

We summarize the contributions of this paper as follows:

- We take the lead paying attention to the continuous immersive assistance for visually impaired people to perceive the world, and first put forward the idea of applying multimodal deep learning to this application scenario.
- We propose a complete system including hardware, data, and algorithms to make the above ideas truly land. We design a flexible tactile sensor glove for the needs of the use scenario. A multimodal attention model is proposed, which can achieve robust classification with high accuracy. We construct a visual-tactile bimodal dataset to train and evaluate our system. For the proposed system, we conduct lots of various evaluation experiments, including the test of the model in extreme situations and the evaluation of the system in real use scenarios.
- We opensource all the code and datasets in hope that researchers can freely use them to promote the development of the field of visual impairment assistance, which will accelerate the practical application of research in this field and bring about a change in the life of visually impaired people.

The remaining of the paper is organized as follows. Sect. 2 surveys the related problem and methodologies. Section 3 detailedly presents the design and implementation of the AviPer system. Section 4 shows extensive experiments we conduct to evaluate the system, including model evaluation and real-world tests. In Sect. 5, we discuss the insights, achievable optimizations, and prospects of our system. Then we conclude in Sect. 6.

2 Related work

The problem and methodologies presented in the paper are highly related to the following three research areas: visually impaired assistance, multimodal learning, and attention mechanism.

2.1 Assistance for visually impaired people

Assistance for people with impairments has always been a hot topic in the field of human–computer interaction and pervasive computing. There have been many works focused on visual assistance, such as Aladren et al. (2014), Praveen and Paily (2013) and Papadopoulos and Goudiras (2005) in navigation and reading accessibility. Deep learning has a broad

application prospect for impairment assistance, especially the assistance for visually impaired people. Paper Poggi and Mattoccia (2016) proposes a wearable mobility aid for visually impaired based on 3D computer vision and machine learning, which achieves effective and real-time obstacle detection. Work Tapu et al. (2017) develops a system called *DEEP-SEE* which realizes joint object detection, tracking, and recognition for visual impairment assistance. Paper Liu et al. (2021) and Delahoz and Labrador (2017) apply deep learning to floor detection. There are also some smartphone-based approaches such as Lin et al. (2017) and APP Seeing AI by Microsoft. However, The above-mentioned existing research, along with Wang et al. (2017), Lakde and Prasad (2015) and Ganz et al. (2014), mainly focuses on navigation and obstacle avoidance. There are other applications like facial recognition for visually impaired people Neto et al. (2016), facilitating search tasks Zhao et al. (2016) and password manager Barbosa et al. (2016). But research on continuous immersive motivational aids for visually impaired people, especially those in educational age, is insufficient. We hope that our proposed system AviPer, which aims to assisting visually impaired people to understand the world, can fill the gap.

2.2 Multimodal learning

Human perception of the world is multimodal. We see objects, hear sounds, smell odors, feel texture, and taste flavors. *Modality* refers to the way in which something happens or is experienced (Baltrušaitis et al. 2018). Unlike machine learning models with a single data source, multimodal machine learning aims to build models that can process and relate information from multiple modalities, which has great potential to provide a stronger understanding ability for the model. Multimodal learning has a wide range of applications. The earliest examples of multimodal learning include audio-visual speech recognition (Yuhas et al. 1989) and multimedia content indexing and retrieval (Snoek and Worring 2005). In the early 2000s, multimodal learning began to be applied to human activity detection (Smith et al. 2005; Yin et al. 2008), as it is inherently very suitable for handling multimodal human behavior. Now, multimodal learning has been widely used in tasks that require a complex perception of the surrounding environment like self-driving (Xiao et al. 2020; Cui et al. 2019) and health monitoring (Banos et al. 2015; De et al. 2015).

The main challenge in multimodal learning is to choose the optimal fusion structure. Deep architectures offer the flexibility of implementing multimodal fusion either as early, intermediate, or late fusion (Ramachandram and Taylor 2017). In early fusion, also known as *data-level fusion*, the various sampling rate of different sensors and huge dimensional differences of heterogeneous data would be tricky. The common approach for alleviating the challenges related to raw data fusion is to extract high-level representations from each modality before fusion, which could be handcrafted features or learned representations, widely used in works like Wu et al. (2016), Karpathy et al. (2014) and Simonyan and Zisserman (2014). Therefore, intermediate fusion is also known as *feature-level fusion*. Late fusion, or *decision-level fusion*, represents a paradigm for fusing the results of network branches handling different modalities. The advantage of this method is that it is feature independent, which means error caused by each modality is uncorrelated. Neural network architecture with intermediate fusion needs careful design. Recently, the use of automatic machine learning to adjust intermediate fusion network architecture has become a hot trend (Ramachandram et al. 2017; Li et al. 2017).

The fusion of tactile and visual perception has also developed for decades. As early as the 2000s, neurologists studied the coordination of vision and touch in human perception (Zangaladze et al. 1999; Ernst and Banks 2002). Work Björkman et al. (2013) and Luo et al. (2015) introduce low-resolution tactile sensing to assist visual tasks. Work Kroemer et al. (2011), Güler et al. (2014) and Gao et al. (2016) respectively propose haptic-visual multimodal deep learning models for specific tasks. The main difference between our work and the above researches is that they focus on robot perception and manipulation while our system uses flexible wearable tactile sensors to increase the user's participation in assistance tasks for visually impaired people, as well as to boost performance.

2.3 Attention mechanism

Attention mechanism is a data processing approach in machine learning. Since first proposed by Bahdanau et al. (2014), it has been extensively used in natural language processing (Hu 2019), computer vision (Sun et al. 2020), and other various machine learning tasks. The main idea of attention mechanism comes from the way humans perceive things, which is expected to put more attention on key features that need more concern. At the implementation level, the basic approach is to use a *mask* to reweight the data, in order to endow the region concerned with higher weights. Attention mechanism can be classified as soft attention, hard attention, and self attention. Soft attention is differentiable, while hard attention is not. Training process of the latter is usually completed through reinforcement learning. Self-attention is a special form of attention mechanism, which focuses on the intrinsic correlation of different elements in the data source, whose representative architecture is *Transformer* (Vaswani et al. 2017) and its variants.

In soft attention mechanism focus on computer vision task, Paper Hu et al. (2018) proposes *Squeeze-and-Excitation*

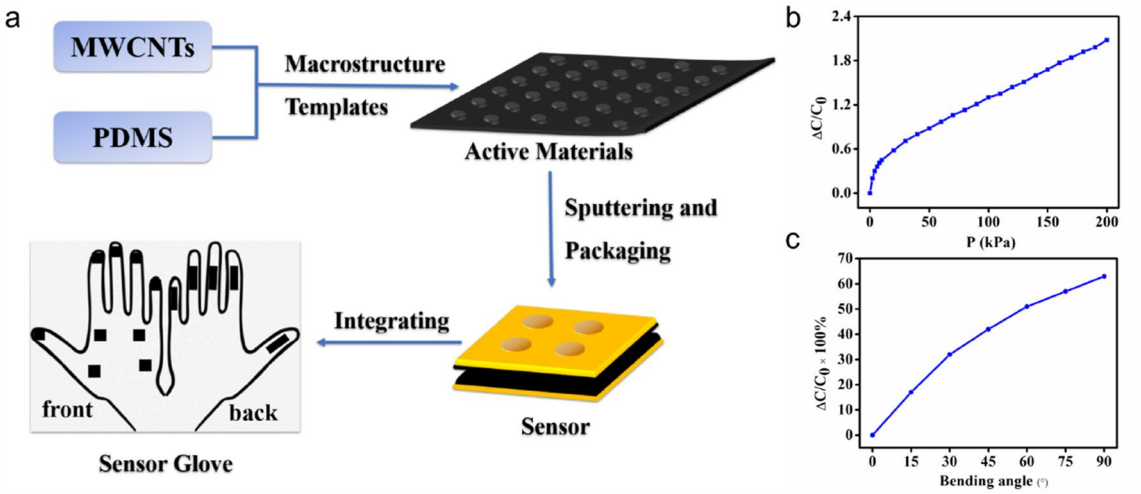


Fig. 2 The tactile glove. **a** Production and integration: MWCNTs (multi-walled carbon nanotubes) powder and PDMS (polydimethylsiloxane) are used to make an active layer film with a regular macroscopic shape. The measurement surface of the thin film structure is placed and packaged face to face and Au nanoparticles are sputtered on the outside as electrodes to produce a flexible capacitive sensing unit. Pressure sensors are placed on the fingertips and palms of the hand, while tension sensors are placed on the joints. **b** The response of the pressure sensors to the force. **c** The response of the tension sensors to the bending angle

Network which carries out channel-wise attention. For spatial attention, Wang et al. (2017) employs the idea of residuals to develop *Residual Attention Network*, which learns attention-aware features from different modules that change adaptively as layers go deeper. Then, paper Woo et al. (2018) develops a channel-spatial integrated attention mechanism. Paper Wang et al. (2018) and its improvement Cao et al. (2019) proposes *Non-local Attention* for CNN architecture, which provides solutions to capture long-range dependencies. For deep learning architecture handling time series data, there are many studies on attention in time-domain like Qin et al. (2017) and Liang et al. (2018) and frequency-domain like Lee et al. (2020). As for research that is highly relevant to our work, Cao et al. (2020) embeds spatial mechanism to tactile sensing for texture recognition. But studies are lacking for attention mechanism in the fusion of tactile and visual modalities. Besides, attention-based research specifically for visual impairment assistance is rather insufficient.

3 System design and implementation

In this section, we will introduce the components of our proposed system in detail. The system consists of three parts: hardware, data, and multimodal attention model. In order to achieve accurate classification of objects, first of all, sensors that collect tactile data and a webcam for recording visual data are required. These hardwares along with the microcontrollers for data transmission, GPU for training, and speaker for broadcasting will be discussed in Sect. 3.1. For training and testing the classification model, we constructed a bimodal dataset and prepossessed the data, whose specific collection strategy and prepossessing pipeline including data augmentation will be given in Sect. 3.2. Then, we detailedly introduce the architecture and settings of the multimodal attention network used for classification in Sect. 3.3.

3.1 Hardware

3.1.1 Tactile glove

To achieve high-precision object recognition while providing users with as much participation and realism as possible, we integrated our self-developed capacitive-based tactile sensor into a wearable glove. By investigating real human grasping processes, we inferred that most of the grasping gestures can be deduced from the degree of bending of the fingers and the force on the fingertips and palms. Hence there is no necessity to use high-density sensors like Sundaram et al. (2019), which not only allows our gloves to be manufactured at a relatively low cost, but also sharply reduces the wiring difficulties and maintenance costs, which is more in line with the actual use requirements.

As shown in Fig. 2 the sensing glove consists of 14 sensors containing 9 pressure sensors and 5 tension sensors. The pressure sensors are fixed on five fingertips and four corners of the palm, and each sensor unit is about 1.2×1.2 cm^2. While tension sensors were fixed to 5 finger joints and each sensing area is about 1×1.5 cm^2. Both pressure and tension sensors are fabricated by two thin films as active layers and Au nanoparticles as electrodes. Changes in shape and

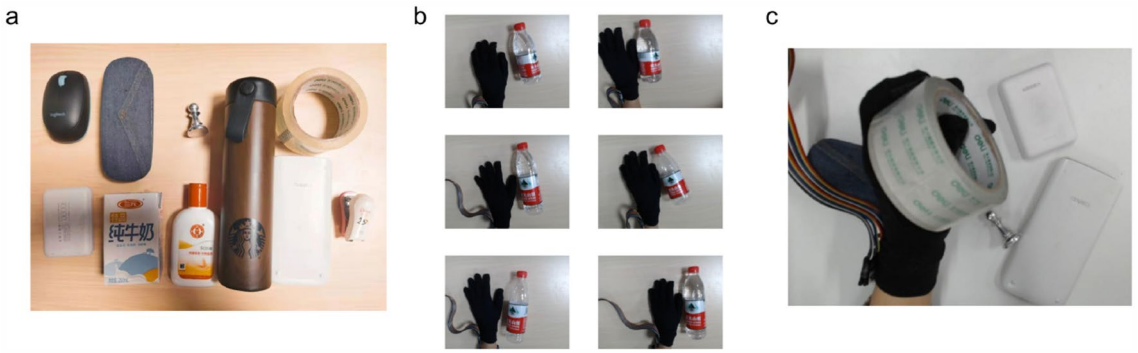


Fig. 3 Data collection. **a** The 10 objects used for constructing the main dataset. **b** When constructing the second dataset, which is made up of objects that are hard to recognize only by vision, we choose two bottles of different sizes. Each one is grasping under the following 3 status: full, half full, empty. In **c**, we collect the same 10 items as **a** but under the condition where there are distracting objects on the table

electrode distance are manifested as changes in capacitance. Therefore, the pressure sensor on the palm and fingertips can give different signals when grasping objects with different forces. And the tension sensors on the joints can show different signals when grasping objects with different shapes as the fingers bend at different angles.

3.1.2 Others

Besides the tactile glove mentioned above, the hardware of our system includes two Arduino Mega microcontrollers, which are used to transmit signals between the tactile sensor and the computer serial port. In addition, we use an easily available web camera (Logitech c922 Pro Stream webcam) to collect video data. The above-mentioned microcontrollers and webcam are directly connected to the host computer through a USB interface. For the multimodal attention model, we complete all training and evaluation experiments on an NVIDIA Geforce GTX 3070 GPU. Finally, we use a Bluetooth speaker (MC A7) to announce the predicted results to visually impaired users.

3.2 Data

3.2.1 Data collection

In the scenario of assisting visually impaired people to learn to recognize objects, we look forward to give prediction results based on a short period of grasping action. The time period should allow users to fully perceive the characteristics of the object, and at the same time allow the model to give an accurate prediction. Therefore, we adopt the strategy of collecting hand motion videos and tactile time series of every complete grasping. Specifically, in the basic classification experiment, we first select ten types of items that are commonly used in life, which are illustrated in Fig. 3. For each type of objects, we have carried out more than 150 times of complete grasping: picking it up, holding it for a few seconds, then putting it down, which forms the original dataset (a). When grasping the same item, we adopt a variety of gestures to enhance the diversity of data from the aspect of data collection, which is tally with the actual use situation as well. Besides, we choose two different desk textures as background. One is warm with wood grain, and the other is cool without grain. The ratio of data in the two backgrounds is about 2:1. Each piece of data includes a grasping video (.avi) with a length of about 17 s, tactile sensor information for the same time period (.csv), and four frames (.jpg) of the 0th, 5th, 10th, and 15th s intercepted from the video.

In order to demonstrate the robustness of visual and tactile multimodal prediction to extreme situations, we design 6 categories of items that are difficult to distinguish only from the camera data. Each category is collected more than 100 samples with the same collection pipeline described above but only on warm grained background, which makes up dataset (b). We build another small dataset (c) which contains 200 pieces of data of 10 categories in total. In the collection of these parts of data, irrelevant objects are placed in the view field, which aims to test whether the participation of tactile information can help the model resist interference and whether spatial attention can focus on the operating hand. We finish this part of collection on the cool no-grained desk. The items that we use to collect data are shown in Fig. 3. We perform data augmentation to enlarge the size of all the three datasets at a ratio of 1:10. Data augmentation and preprocessing will be discussed in Sect. 3.2.2. Details of the augmented dataset configurations can be found in Table 1.

3.2.2 Data augmentation and preprocessing

Data augmentation is a data-space solution to the problem of limited data (Shorten and Khoshgoftaar 2019). By flips,

Table 1 Dataset configurations

Dataset	Categories	Size	Background	Ground truth labels
(a)	10	16,500	Warm:cold ≈ 2:1	Book sewer, calculator,chess, glasses box, vacuum cup, hand cream, milk, mouse, power bank, tap
(b)	6	6600	Cold	Empty large, empty small, half large, half small, full large, full small
(c)	10	2200	Cold	Book sewer, calculator,chess, glasses box, vacuum cup, hand cream, milk, mouse, power bank, tap

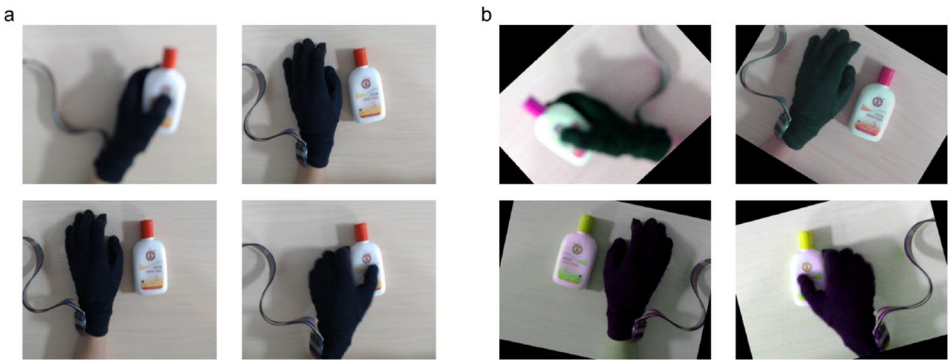


Fig. 4 Data augmentation example: *hand cream*. **a** shows the original video frames of different time points. **b** Shows the 4 frames after data augmentation

translations, and rotations, we could significantly expand the dataset. For tactile time series, there are argumentation approaches in time domain and frequency domain (Wen et al. 2020). In our experiments, we simply augmented visual data while keeping tactile time series unchanged. We generated new video frames from the collected datasets (a), (b), and (c) at a ratio of 1:10 following the process:

1. Color jitter: adjusting brightness, contrast, saturation, hue ±20%
2. Random rotation in interval [−45°, +45°]
3. Random vertical flip and horizontal flip both with a probability of 0.2
4. Random gray scale with a probability of 0.2

It is worth mentioning that we implement data augmentation for 4 frames from the same video independently, that is each of the 4 frames go through different data augmentation configurations, which enhances the diversity of the augmented datasets. An argumentation example can be seen in Fig. 4.

The function of data preprocessing component is to generate the data which can be directly fed to the multimodal deep learning model. Due to the characteristics of the capacitance-based tactile sensor, the initial value (which refers to the value of the natural placement state without grasping anything) and range of change of the raw tactile data vary in different tactile sensor units, which is shown in Fig. 5. The huge difference in raw tactile time series among each channel will cause weight bias of each channel's importance to the trained model. Therefore, in order to let the model learn the information of each unit more comprehensively and evenly, we independently normalize each tactile sensor channel. Specifically, we find out the maximum and minimum values of all the data of a certain sensor unit, denoted as $V^{(i)}_{\max}$, $V^{(i)}_{\min}$, then use the following simple linear mapping to map all the time series recorded by the sensor to [0, 1]:

$$V'_i(t) = \frac{1}{V^{(i)}_{\max} - V^{(i)}_{\min}} V_i(t) - \frac{V^{(i)}_{\min}}{V^{(i)}_{\max} - V^{(i)}_{\min}}, \quad i = 0, 1, \ldots, 13.$$

For the four frames of images used for training, we also adopt normalization. First, we convert the RGB image to a tensor in the range of [0, 1]. Then, the mean is adjusted to 0 and the standard deviation to 1. The data prepossessing pipeline is illustrated in Fig. 5.

3.3 Visual and tactile multimodal attention network

In this section, we will discuss the proposed multimodal attention network to classify objects with tactile and vision data in detail. Concretely, as the data we collect are time series of 14 tactile sensor units all over the glove and 4 frames taken from the particular time in the corresponding downward videos of hand, we carefully design 4 distribution functions p_i $(i = 0, 1, 2, 3)$ to hard-code the importance of tactile signals at different times to make tactile-map and video image pairs, denoted by $(t_i, v_i), (i = 0, 1, 2, 3)$. On account of only a portion of sensor units play a major role in the process of grasping, to fully extract the

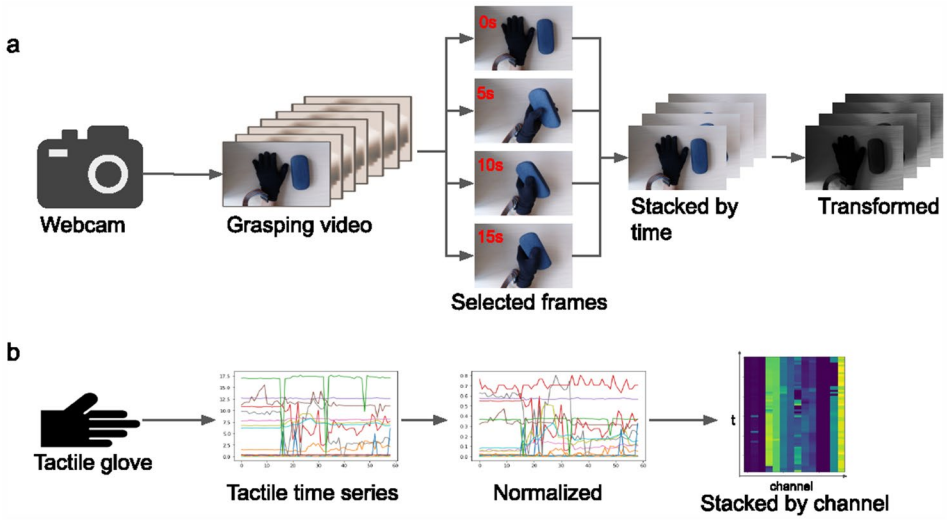


Fig. 5 Data preprocessing: **a** shows the preprocessing pipeline of video data. We first select 4 special frames from the original video and then stack them by time and implement transform to normalize them. For tactile time series, as shown in **b**, we first implement the normalization and then stack them by sensor channels

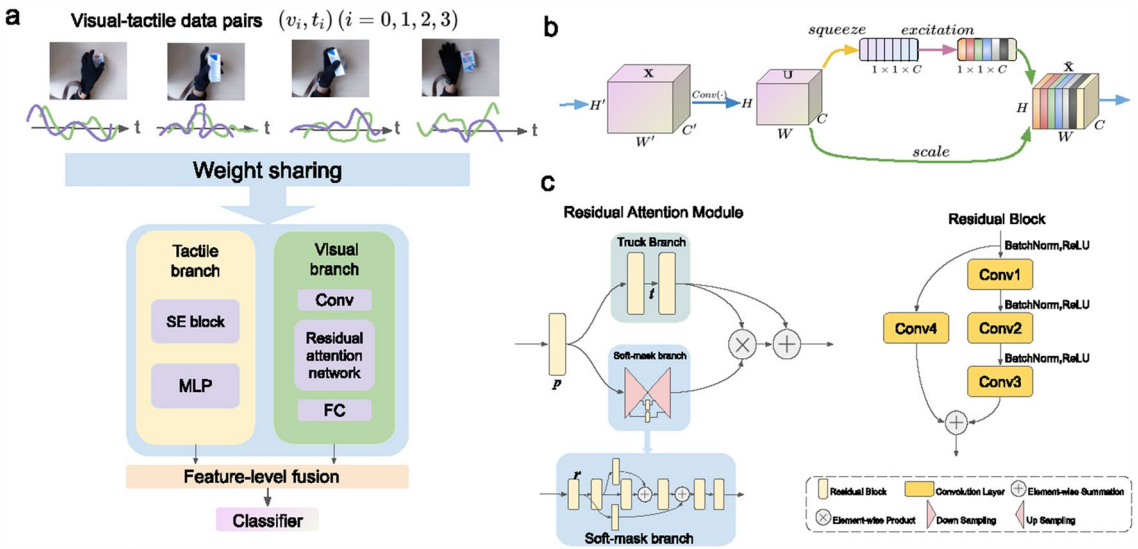


Fig. 6 Network architecture. **a** Shows the overall structure of the model, in which we adopt weight sharing strategy to learn from different visual-tactile data pairs $(v_i, t_i), (i = 0, 1, 2, 3)$ of a single grasping process. The network is divided into tactile and visual branches, based on *squeeze-and-excitation network* (Hu et al. 2018) and *residual attention network* (Wang et al. 2017) respectively. The structure of SE blocks is shown in **b**. $\bar{\mathbf{X}}$ is the SE block output of original feature $U = Conv(\mathbf{X})$ as input. The structure of a single residual attention module is shown in **c**. In our residual attention network, we stack 3 residual attention modules along with max-pooling layers and convolution layers

information of different tactile units, we embed squeeze-and-excitation(SE) blocks in the branch of the network that extracts haptic information, which achieves channel-wise attention. For a single video frame, we apply attention to spatial position through residual attention modules to locate hand area to make the prediction more robust. We extract feature vectors from data by visual network branch and tactile network branch respectively, then perform a feature-level modal fusion with a modality weight hyper-parameter λ. Ultimately, we pass the fused features with relatively low dimensions to a two-layer perceptron to give the prediction. It is worth mentioning that although we have 4 different (t_i, v_i) pairs, each of them will be put into two branches(tactile branch and vision branch) of exactly the same network, which means the weights are shared when the proposed model process data pairs from disparate temporal interval. The overall schematic diagram is shown in Fig. 6, the structural details of separate parts of the proposed network architecture will be discussed in the following subsections.

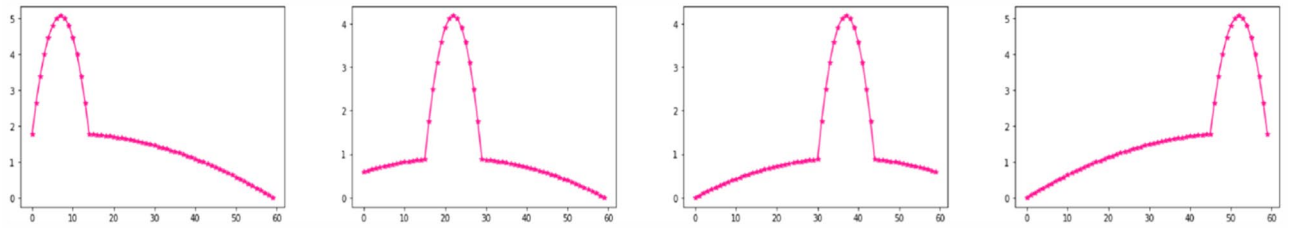


Fig. 7 Weight distributions, from left to right labeled as No. 0, No. 1, No. 2, No. 3

3.3.1 Hard-coded temporal attention

As mentioned in Sect. 3.2.1, the vision data we use is 4 single frames cut out from the full-length video of grasping objects rather than the video itself which requires lots of computational resources. The chosen 0th, 150th, 300th, 450th frames of the video lasting about 15 s represent the hand gesture of a specific temporal interval. It would be natural to think that we should also weigh tactile time series in time dimension to distinguish the importance of the current frame time interval from other intervals. Therefore, we design 4 quadratic distribution function to depict attention in time. As the tactile series' length is 60, we divide it to four intervals: $0 \sim 14, 15 \sim 29, 30 \sim 44, 45 \sim 59$, denoted by time period No. 0, No. 1, No. 2, No. 3. During each time period i, the starting index is denoted as T_i^s, the ending index is denoted as T_i^e, The distribution function $p_i, (i = 0, 1, 2, 3)$ is given by the following equation.

$$p_i(t) = \frac{f_i(t) + g_i(t)}{s} \tag{1}$$

s is a scale factor which is set to 15. While, $f_i(t)$ is the peak distribution, which is defined as:

$$f_i(t) = -(t - T_i^s) \cdot (t - T_i^e), \quad t \in [T_i^s, T_i^e]. \tag{2}$$

$g_i(t)$ is the body distribution, which is defined as:

$$g_i(t) = \begin{cases} -w \cdot (t - 59) \cdot (t - T_i^s - T_i^e + 59), & i = 0, 1 \\ -w \cdot t \cdot (t - T_i^s - T_i^e), & i = 2, 3 \end{cases}, \quad t \in [0, 59], \tag{3}$$

where w is the weight to adjust the ratio of the maximum of $g_i(t)$ and $f_i(t)$. Outside their domain of definition, both functions are set to 0. The weight distribution obtained is shown in Fig. 7. For each distribution, we sample at the integer index between 0 and 59 to form a 60-dimensional vector $\mathbf{p_i}$, then the weights are hardcoded into tactile time series of each sensor unit j by element-by-element multiplication. So that we end up with 4 tactile-visual data pairs $(v_i, t_i), i = 0, 1, 2, 3$ which imply 4 adjacent time intervals.

$$t_i^{(j)} \leftarrow t_i^{(j)} \odot \mathbf{p_i}, \quad j = 0, 1, \ldots, 13 \tag{4}$$

3.3.2 channel-wise tactile attention with squeeze–excitation blocks

By reason of that the tactile sensors are all over the palm and knuckles, it is very unlikely that each sensor has a strong and similar signal change in a single grasping action. To exploit sensor channels' different importance dynamically, we embed a SE block in tactile branch of the model, which is proposed by Hu et al. (2018), containing two processes, namely *squeeze* and *excitation*, shown in Fig. 6b. In the *squeeze* process, we generate channel descriptor $\mathbf{z_i} \in \mathbb{R}^{14}, (i = 0, 1, 2, 3)$ using global average pooling:

$$z_i^{(j)} = \frac{1}{60} \sum_{m=0}^{59} (t_i^{(j)})_m, \quad j = 0, 1, \ldots, 13. \tag{5}$$

Following comes the *excitation* process to make use of the information integrated into *squeeze* process, which is implemented with a gating mechanism:

$$\mathbf{s_i} = \sigma(W_2 \delta(W_1 \mathbf{z_i})), \quad i = 0, 1, 2, 3, \tag{6}$$

where δ refers to the ReLU activate function, σ refers to a Sigmoid function, $W_1 \in \mathbb{R}^{\frac{14}{r} \times 14}$, $W_2 \in \mathbb{R}^{14 \times \frac{14}{r}}$. r is the reduction ratio, which was set to 2 in our experiments. At the end of *excitation*, the original input t_i is rescaled by channel:

$$t_i^{(j)} \leftarrow \mathbf{s_i}^{(j)} t_i^{(j)}, \quad j = 0, 1, \ldots, 13. \tag{7}$$

3.3.3 Spatial visual attention with residual attention modules

In our application scenario, although the gloved hand only grips one item at a time, there is no guarantee that no interferential objects are on the desktop within the camera view. To prevent other objects on the table from interfering with the inference of the visual branch of the model, we propose to make the model capable of perceiving spatial attention to locate the object being grasped. We adopt the approach of Residual Attention Module proposed in Wang et al. (2017), which gets ideas from residual learning and adds soft mask attention mechanism on an identical mapping:

$$H_{i,c}(x) = (I + M_{i,c}(x)) \odot F_{i,c}(x) \tag{8}$$

where i ranges over all spatial positions and c is the index of image channels. $M_{i,c}(x)$ ranges from [0, 1] and, with $M_{i,c}(x)$ approximating 0, $H_{i,c}(x)$ will approximate original features $F_{i,c}(x)$. Meanwhile, the identical mapping guarantees that its performance will at least be no worse than original no-attention network.

The structure of a single residual attention module is shown in Fig. 6c, whose main components are the truck branch and the mask branch. In the construction process, there are 3 hyperparameters p, t, r to control the block's size, respectively denoting the number of preprocessing Residual Units before splitting into trunk branch and mask branch, the number of Residual Units in trunk branch, and the number of Residual Units between adjacent pooling layer in the mask branch. In our implementation, the set values are $\{p = 1, t = 2, r = 1\}$. We stack 3 residual attention modules along with max-pooling layers and convolution layers to form the visual residual attention network. The corresponding size of the 3 residual attention module's outputs are $56 \times 56@256$, $28 \times 28@512$, $14 \times 14@1024$.

3.3.4 Weight sharing

As mentioned in Sect. 3.3, although there are 4 different $(t_i, v_i), (i = 0, 1, 2, 3)$ pairs for network input, but we use a single network with visual branch $V(x)$ and tactile branch $T(x)$ rather than four separate ones, which means the weights are shared when the proposed model processes data pairs from disparate temporal interval. Therefore, each time we update the weights, the gradient comes equally from the four data pairs. This not only greatly reduces the number of network parameters, but more importantly, it indirectly provides more data for single network training. The desired visual branch should have the ability to extract features at any moment, so should the tactile branch. Through weight sharing, we can train network branches with stronger representational ability.

3.3.5 Modal fusion with tunable weight λ

The two branches' outputs $V(v_i)$ and $T(t_i)$ will have exactly the same dimension, denoted as fusion dimension D_f, which is also a hyperparameter (set to 100 in our experiments). We use another parameter $\lambda \in [0, 1]$ to indicate the importance of each modality while performing modal fusion. We set the weight of visual features to λ and tactile features to $1 - \lambda$. In our preliminary trial, λ is set to 0.5, which means that visual and tactile features are equally important in this case. The fused feature vector $\mathbf{v}$ is given by:

$$\mathbf{v} = \sum_{i=0}^{3} [\lambda V(v_i) + (1 - \lambda) T(t_i)] \tag{9}$$

3.3.6 Classifier

The fused feature vector $\mathbf{v}$ will be passed through a two-layer perceptron. The output dimension is the number of object classes. We set the dimension of the hidden layer to $\lfloor \sqrt{D_f + \frac{1}{2}} \rfloor$. With hidden layers of reasonable size, the two-layer classifier provide stronger nonlinearity for the model thus helping to achieve higher classification accuracy.

4 Experiments

In this section, we first detailedly discuss experimental settings in Sect. 4.1, then provide the complete evaluation result in Sect. 4.3 and ablation study in Sect. 4.3. In Sect. 4.4, the robust classification results of the system in several extreme cases are presented. Then we discuss the effects of input data pair selection in Sect. 4.5. Finally, we provide the actual use scenario test in Sect. 4.6.

4.1 Experimental settings

We will clarify the general settings in our evaluation experiments in the first place. For the basic ablation studies in Sect. 4.3, we split the multimodal dataset for training and testing at a ratio of 7:3. During the training process, Kingma et al. (2014) algorithm is employed to optimize the learnable model parameters. We set the learning rate to 0.0005 and the batch size to 4. We choose Cross-entropy Loss as the loss function and use the accuracy score as our performance metric. All of our model implements and experiments are based on Pytorch Paszke et al. (2019). The training and testing processes are completed on a Windows10 PC with an NVIDIA Geforce GTX 3070 GPU. Other specific settings will be given in the corresponding section of the experiment.

4.2 Best accuracy performance

With the 3 kinds of attention mechanism and 2 kinds of modality, our system achieved **99.75**% classification accuracy, which is about 1.5% higher than result of similar network structures with roughly the same number of parameters but no attention mechanism and significantly higher than the model with single modality($\sim$ 5% of single tactile model and $\sim$ 2% of single visual model).

Table 2 Ablation study on attention mechanisms

Attention experiments(A−)	Tactile		Visual	Accuracy/%
	Temporal	Channel-wise	Spatial	
TCS	✓	✓	✓	**99.75**
TC	✓	✓	○	98.71
TS	✓	○	✓	99.14
CS	○	✓	✓	99.02
S	○	○	✓	98.97
T	✓	○	○	98.62
C	○	✓	○	98.77
All moved	○	○	○	98.28
Modality experiments(M−)	Tactile	Visual	Accuracy/%	
TV	✓	✓	**99.75**	
T	✓	×	94.54	
V	×	✓	97.98	

Modality experiments(M-)	Tactile	Visual	Accuracy/%
TV	✓	✓	**99.75**
T	✓	×	94.54
V	×	✓	97.98

Attention experiments compare the classification accuracy /w or w/o attention mechanisms of different kinds: hard-coded temporal attention, channel-wise attention with SE blocks, visual spatial attention with residual attention modules. *Modality* experiments compare the accuracy when different modality is used. In both the two parts of experiments, the used attention mechanisms or modalities are denoted by ✓, while the abandoned attention mechanisms are denoted by ○ in *Attention* set of experiments and the abandoned modalities are denoted by × in *Modality* set of experiments

Bold indicates best performance

4.3 Ablation study

The ablation study of the multimodal attention network on dataset(a) shown in Table 2 is intended to show the necessity of each component. All the experiments done in this section follow the setting in Sect. 4.1.

It's worth mentioning that we abandon the specific attention mechanism by means of setting all of its parameters to 1 if they are attention weights or 0 if they are soft-masks and remove modality by resetting the modality fusion hyper-parameter λ to 0 or 1, respectively representing removing visual and tactile modality. From the experiment results, we can draw a conclusion that the modules we integrate distinctly improve the accuracy of the model. Concretely, with both visual and tactile modalities and all of the three attention mechanisms loaded, the model can achieve an extremely high accuracy of **99.75%**. If we abandon the visual spatial attention, then the test accuracy drops down to 98.71%, shown in A-TC. If we remove the channel-wise attention and temporal attention respectively, the accuracy correspondingly descends to 99.14% and 99.02%, shown in A-TS&CS, whose decreases are less than A-TC. We can infer that visual attention matters more than those two kinds of tactile attention mechanisms. In A-S experiment of ablation study, we remove both of the two kinds of attention mechanisms applied on tactile modality. The accuracy is 98.97%, still better than 98.71% in A-TC experiment, which is in line with our inference above. If we remove all the attention mechanisms, ending up with an accuracy of 98.28%, which implies the effectiveness of introducing attention mechanisms to the classification process. As for ablation experiments for *Modality*, we remove the visual branch of the network and only use tactile information, the accuracy drops sharply decrease to 94.54%, shown in M-T. And the accuracy is 97.98% if only inferring with visual modalities, namely M-V. These two experiments together prove that the modal fusion is rewarding.

4.4 Multimodal prediction in extreme cases

4.4.1 Classification of almost visually indistinguishable items.

In this section, we test the classification modal in the case that items are hard to distinguish only by vision, in order to prove that the integration of tactile modality improves the robustness of the classification model. It can be expected that there are some items that are untoward to distinguish with the webcam alone. However, these items usually have very different tactile characteristics, such

Table 3 Multimodal prediction in extreme cases (*I*): classification of almost visually indistinguishable items

Data modality		Accuracy/% on 6 classes	Accuracy/% on 3 filling states	Recall/% on 2 sizes
Tactile	Visual			
×	✓	93.71	94.06	97.70
✓	×	92.45	94.51	97.74
✓	✓	**99.15**	**99.47**	**99.81**

The used and abandoned data modality are denoted by ✓ and ×. *Accuracy* on 6 classes and 3 filling states and *Recall* on 2 sizes are used as the performance measure

Bold indicates best performance

as weight, texture, material, feeling, etc. Based on the above features, we ingeniously selected 6 types of bottle as targets to be identified, which are even indistinguishable by a human from a little far away. Specifically, we made use of two bottles with exactly the same color but significantly different volume (380 mL/550 mL). Each size bottle was used to collect three data sets with three different filling states: empty, half full, and full, as shown in Fig. 3. We conducted classification experiments on the dataset(b) of above 6 kinds of bottles. We tested three cases separately: using visual data only, using tactile data only, and visual-tactile modal fusion. Results showed that relying on the auxiliary recognition of tactile information can significantly improve the results of visual classification. For specific results, see Table 3. Besides the *accuracy* of classification on 6 bottles, we evaluated the classification *accuracy* for the 3 water-filled states. We also calculated the *recall* rates of different size bottles (by marking the large bottle as positive), which derives from:

$$Recall = \frac{TP}{TP + FN}, \quad TP \text{ for true positive and } FN \text{ for false negative.} \tag{10}$$

It can be seen that the classification result combining the two modalities significantly improved compared to making predictions with data with single modality. The experiment also proves that our proposed system has the potential to achieve more than just object classification assistance. The above example of accurately distinguishing bottles with different capacities and different filling states shows that the system can qualitatively assist visually impaired people to perceive the size and weight of items, which means to let them feel the characteristic of items.

There is also a large category of items that are easy to distinguish visually but almost the same in touch, for example, the items of the same shape but different colors. But this kind of object is of little practical significance to our application of assisting visually impaired people.

Table 4 Multimodal prediction in extreme cases (*II*): classification when there are interfering objects on the desktop

Mixing proportion				Accuracy/%
Dataset(a)		Dataset(c)		
Train/%	Test/%	Train/%	Test/%	
70	30	70	30	94.77
90	10	50	50	92.54
70	30	30	70	92.47
100	0	50	50	89.97
50	50	100	0	88.55
0	100	100	0	52.78
100	0	0	100	49.96

Different mixing proportion for dataset(a) and (c) to form the training set and test set is evaluated and compared

Additionally, the advantages of visual recognition have been proved by a large number of previous studies. Therefore, we did not implement these additional experiments.

4.4.2 Classification when there are interfering objects on the desktop.

In this section, we test the classification model in the case of interfering objects placed messily on the table. In practical situations, it is common for more than one item to appear in the field of view captured by the webcam, while only one of these objects is the target that the visually impaired user is grasping, which is the one model should predict. Specifically, we mixed dataset(c) of 10 items mentioned in Sect. 3.2.1, approximately 2200 pieces of data after augmentation in total, and the original dataset(a) in varying proportions to form the training set and test set. During each grasping process in the collecting process of dataset(c), there is more than one item in the view field. We labeled the data according to the items which are directly grasped by the gloved hand. The results evaluated on the mixed dataset are shown in Table 4.

From the results, we can see that in the first five experiments, the classification accuracy is considerably high even if the proportion of each dataset in the training set or the test set varies. Concretely, in the first five experiments, data from each of the two datasets are partly used to form the training set, which ensures the capability of the model to learn essential features of these 10 items and robust attention mechanism to overcome the disturbance of irrelevant objects. However in the other two experiments, when we separately use one dataset to train and the other to evaluate, the accuracy declines steeply. The conceivable underlying cause is that the visual attention training is misled. Therefore we tried removing visual branches to see the impact of visual misleading. The classification results

Table 5 Input frame selection

Indexes of data pairs	Accuracy/%	Indexes of data pairs	Accuracy/%
w/ hard-coded temporal attention			
0, 1, 2	99.12	1, 2	98.98
0, 1, 3	99.31	1, 3	98.79
0, 1, 4	99.17	2, 3	98.85
1, 2, 3	98.82	0	98.04
0, 1	98.57	1	97.93
0, 2	98.94	2	97.79
0, 3	99.06	3	98.15
w/o hard-coded temporal attention			
0, 1, 2	98.54	1, 2	98.07
0, 1, 3	98.67	1, 3	97.94
0, 1, 4	98.80	2, 3	98.28
1, 2, 3	98.19	0	97.89
0, 1	98.07	1	97.73
0, 2	98.24	2	97.91
0, 3	98.23	3	97.84

The experiment was done on dataset (a) with various combinations of data pairs except for complete combination which is already evaluated in Table.2(1). To see the role of the hard-coded temporal attention mechanism in fully extracting features of data pairs in different time intervals, we did the same test under two conditions: w/ or w/o hard-coded temporal attention

of the last two experiments with single tactile modality are **84.43%** and **88.57%** respectively, which proves that ill-trained visual branch does severely adversely affect the results.

4.5 Effects of input data pairs (v_i, t_i) selection

The main purpose of this experiment is to explore the influence of the number of input tactile-visual data pairs (v_i, t_i) on classification accuracy. In the previous experiments, we found a difference between the collected data and the actual use scene. Due to every piece of data in the dataset records an entire grasping process. Therefore, the first frame of each group of video frames always starts with the user's hand next to the object to be classified, while in the actual use scene it is possible that the object to be identified is held or hidden in gloves during the entire identification process. In order to verify the impact of input data pairs selection on the experimental results, we used different input data pair combinations to perform classification experiments. Details are shown in Table 5.

The results show that for the same number of input frames, there is no significant difference in classification accuracy. But reducing the number of input frames will result in a slight decrease in classification accuracy, compared to our best accuracy of 99.59%. Therefore, the frame selection strategy we use should not have a major impact on the actual use scene. In addition, the accuracy in experiments done without hard-coded temporal attention is less than ones with hard-coded temporal attention, which reflects that the proposed temporal attention effectively helps the model learn features from data pairs of different time intervals.

4.6 Evaluation in real use scenario

In this section, we implement our system into the actual use scene. Due to the COVID-19 pandemic, we were limited to finding a well-matched and adequate target audience. In line with the principle of epidemic prevention and control, we did not recruit social experimental users but carried out the practical application scenario test among 5 colleagues from the department. However, we believe that this choice has little influence on the validity of our user experiment since our purpose is to test the adaptability of the system to different individual user behaviors. Therefore, the core is to test the predictive performance of the model on data collected by different people in real-time. Specifically, we recruited 5 test users including 3 females and 2 males with normal visions. In order to more realistically simulate the use of the system by visually impaired people, we put on blindfolds for the users during the test, and the users explored the use of the system by themselves. The use scenario is shown in Fig. 8. The 5 users conducted a total of 200 tests, 10 types of items for 20 times each. The number of correct predictions for each category is listed in Table 6, from which we can learn that the accuracy in the actual use scene only drops slightly than which in the test set.

Since the system is designed for continuous assistance, in actual situations, users can strengthen their perceptual knowledge of objects by repeatedly identifying the

Table 6 The accuracy of each item in use scenario evaluation

Category	Vacuum cup	Tape	Glasses box	Power bank	Mouse
Acc.	100%	100%	95%	90%	90%
Category	Calculator	Hand cream	Milk	Book sewer	Chess
Acc.	85%	80%	80%	70%	70%

Total *accuracy*= 86%

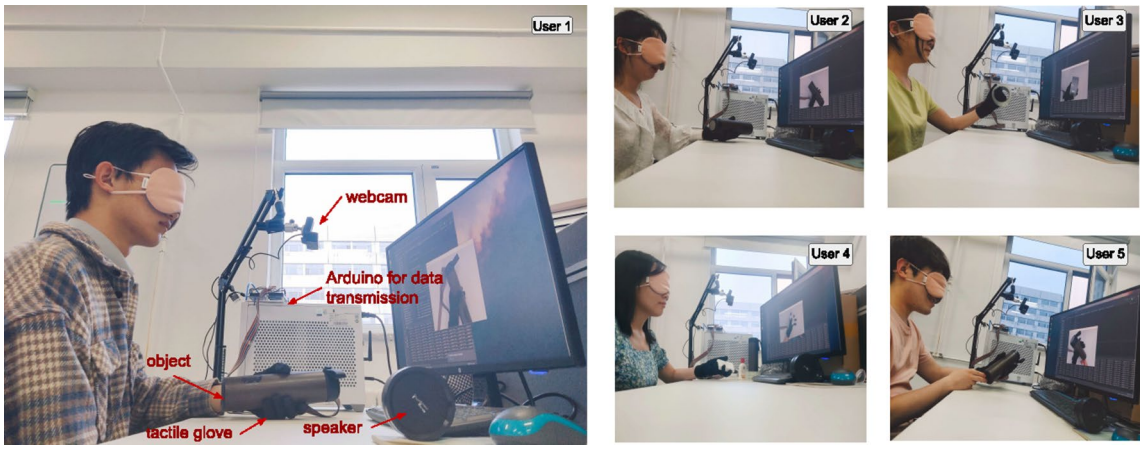


Fig. 8 Actual use scene. The main hardware of our system is marked in User 1's picture. When the mouse is clicked, the system begins to work. The webcam and glove are turned on at the same time to collect data. The blindfolded user then repeatedly changes the posture of his gloved hand to grasp the current object to build his own cognition of the item and give his own judgment. When the collection is done, the multimodal classification model quickly broadcasts the predicted label via the speaker. Because the user's purpose is to continuously learn to distinguish the objects in his gloved hand through the process, if the user is in doubt about the result, he can repeat the process, change the gesture, and then listen to the prediction result again, eventually form the cognition of the item in hand

same object, the accuracy of the actual application will be higher than the accuracy of simple classification, because there are repeated recognition in most instances. Besides, there is an obvious trend that large-size and obvious-shaped items have better prediction results, while small and light items are predicted poorly, which is in line with the result on the test set.

The user-friendliness of the system is mainly reflected in the fact that for users who are blindfolded, after a brief oral introduction in advance, all of them can master the use of the system within a few minutes. This proves the potential of our system and that it would be easy to use for serving truly visually impaired people as well. From the feedback of the test-takers, the five people were impressed with the actual operation feeling after putting on the gloves. In addition, everyone has realized the process of repeatedly grasping, predicting, and finally determining what the object is in the hand. In our opinion, it can be used by visually impaired people equally easily. Whenever there is an opportunity, we hope to recruit visually impaired users to do more in-depth usage tests.

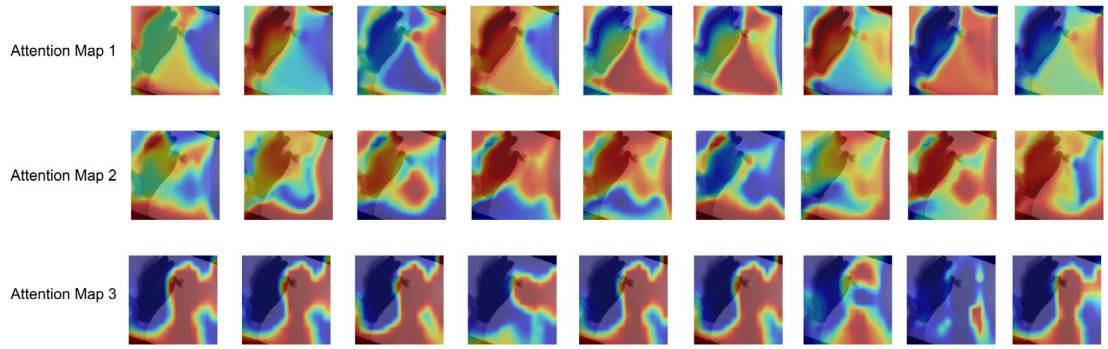
5 Discussion

5.1 Visualization of spatial and channel attention

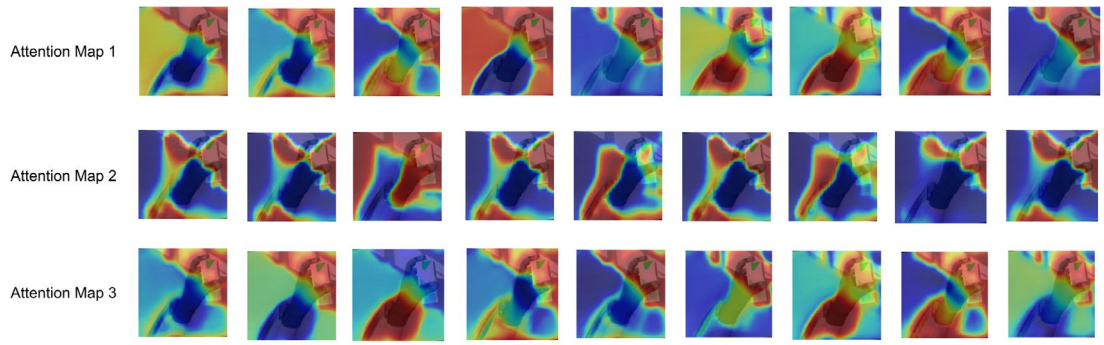
In this section, we made an attempt to visualize the learned spatial and channel attention weights so as to examine the effectiveness of attention mechanisms and give more insights.

5.1.1 Spatial attention

As mentioned in Sect. 3.3.3, the vision branch of our model has 3 residual attention blocks to enforce spatial attention perception. The size of the attention map is $56 \times 56@256$, $28 \times 28@512$, $14 \times 14@1024$ correspondingly, which can provide attention focus from local region to global features. We visualized every channel of the attention maps calculated by these 3 attention blocks, which shows excellent effect in focusing on the tactile glove and the grasped object. Some of the visualizations are shown in Figs. 9, 10, and 11, in which we randomly sampled several channels of the three attention layers. From this, we can preliminary see how the attention mechanism works to improve the performance of the model. With the layer of attention network getting deeper, attention mechanism tries to focus gradually on the area of interest. Note that not every channel of the particular attention layer can ensure to pay attention to valid regions. But as the number of attention map channels is large, which statistically guarantees that attention mechanism works effectively. Especially, in the visualization on dataset(c), in which interfering objects appear in view, attention mechanism helps the model focus correctly on the grasped objects, shown in Fig. 11. Comparing the visualization results of attention maps obtained from the experiments on the three data sets, we can see its superiority in focusing on key regions and eliminating interference features.



(a) Attention map samples of class *chess*.



(b) Attention map samples of class *milk*.

Fig. 9 Visualized attention maps on dataset(a). Objects of small size rely more on attention maps in the deeper layers, while those of larger size have effective attention at both shallow and deep levels

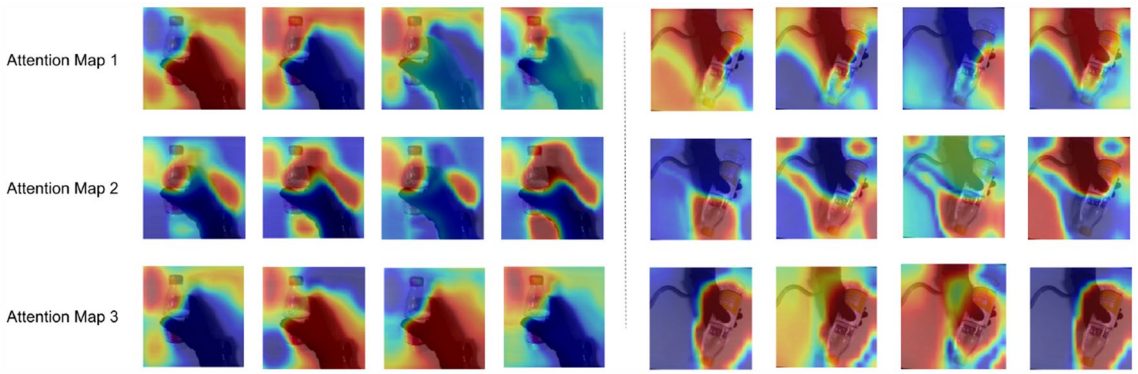


Fig. 10 Visualized attention maps on dataset(b). Left: *empty-small bottle*. Right: *full-large bottle*. Spatial attention is still very effective, but the visual appearance between classes is too similar to provide distinguishing features

5.1.2 Channel attention

In Sect. 3.3.2, we detailed introduced the channel attention in our proposed model. It aims to select several sensor channels with more importance. In practice, we found that a well-trained channel-wise attention module will give different sensor channels significantly different weights, shown in Fig. 12, which verified its validity. Besides, we observed that inputs from different categories only change the channel-wise weights slightly, which means channel weights are almost consistent for different object classes.

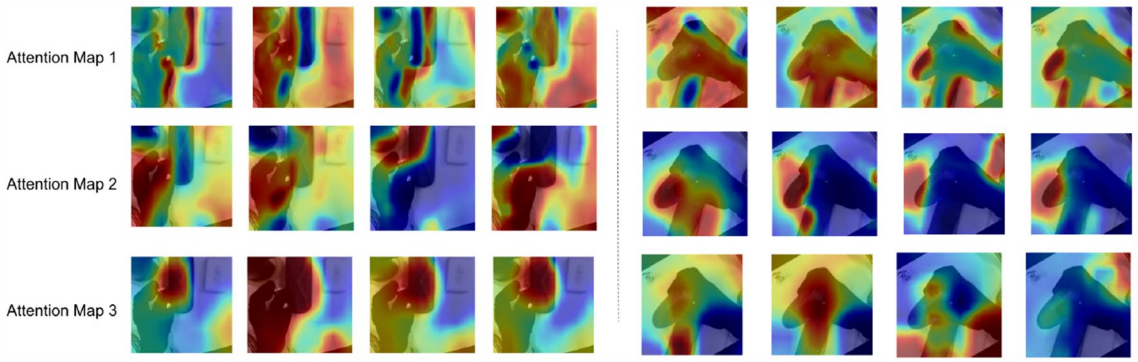


Fig. 11 Visualized attention maps on dataset(c). Left: *chess*. Right: *glasses box*. Attention mechanism helps to highlight objects of interest and suppress irrelevant ones

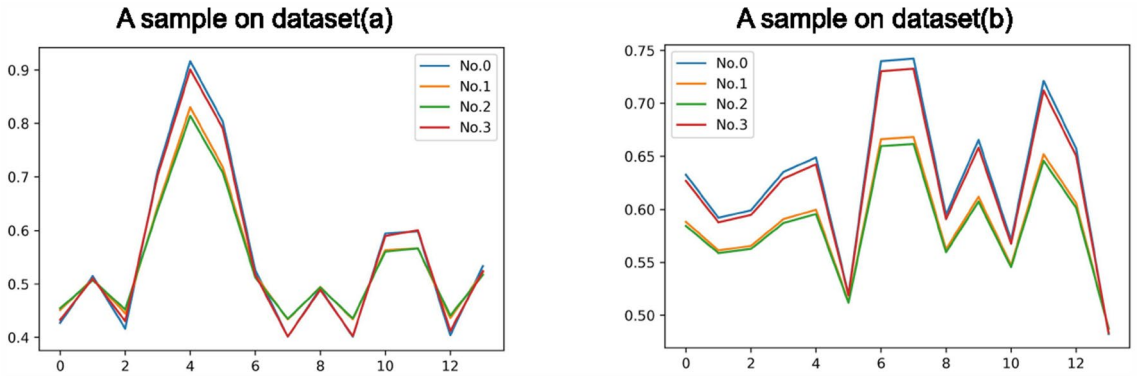


Fig. 12 Channel weights visualization on dataset(a) & (b). The 4 curves in each figure represent channel attention weights of 14 sensors obtained with tactile time series under 4 kinds of time attention as input

5.2 Learnable modality fusion parameter λ

In all of the above experiments, we treated visual and tactile modality equally in the modal fusion process of our model, which means we reckoned without the preference of different objects for one of the two modalities. However, as mentioned in Sect. 4.4, there exist cases in which objects are more likely to be identified if attaching more dependencies to one particular modality due to its distinguishing features of this modality. In order to see the modality preference of the experimental objects and furthermore, use the preference to improve classification accuracy, we adjust the fusion hyperparameter λ to be learnable.

5.2.1 Learning process of λ

Technically, we implemented a tiny network to give the optimal λ for each input data based on the fusion-layer features. Given vision-branch output $V(v_i)$ and tactile-branch output $T(t_i)$ for $i = 0, 1, 2, 3$. Then λ is given by:

$$\lambda = \mathscr{G}\left(\sum_{i=0}^{3} T(t_i) + \sum_{i=0}^{3} V(v_i)\right) \quad T(t_i), V(v_i) \in \mathbb{R}^{D_f}, \tag{11}$$

where $\mathscr{G}$ is a linear mapping followed by a *Sigmoid* function to rein λ in (0, 1). Then, the fused feature is computed by Eq. (9) in which λ weights visual features and $1 - \lambda$ weights tactile features.

5.2.2 Evaluation on 3 datasets

We performed an evaluation on all of the three datasets (a), (b), and (c). In the basic dataset (a) with 10 classes, we randomly sampled 100 times from each category and calculate the fusion factor λ and $1 - \lambda$ of each sample from the trained model. The results showed that in most categories, visual modality is clearly preferred, the λ varies from 0.6 to nearly 1.0, but in three categories, namely calculator, hand cream, and tape, λ is rather low, which indicates that tactile

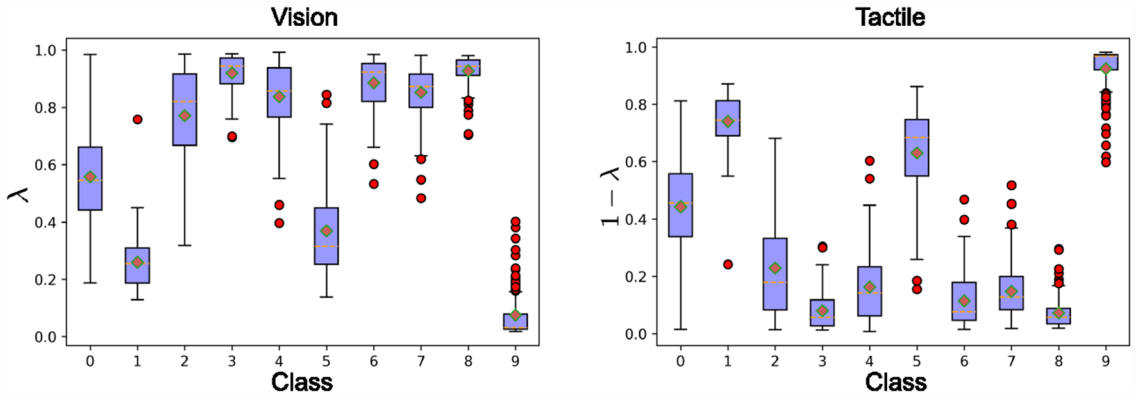


Fig. 13 The fusion factors λ and $1-\lambda$ for experiments on dataset(a). See Table 1 for category labels. The yellow dotted line is the median line and the orange diamond is the mean point. The red points are extreme outliers

dominates in the classification of these three classes. The box plots (Figs. 4a, b, 13) give the details of visual and tactile fusion factors of each class.

In the experiments done on dataset (b) and (c), significant results appeared. Both of the two experiments learned a λ close to 0 for all categories in the datasets, which means that with the self-adapting fusion factor, the model tended to reference using tactile features almost exclusively while ignoring visual features. As mentioned in Sect. 4.4, both of these two experiments represent a kind of visual indiscernibility. The learned fusion factor reasonably confirms the dominant position of tactile features in these cases, which highlights the necessity of introducing tactile modality, shown in Fig. 14.

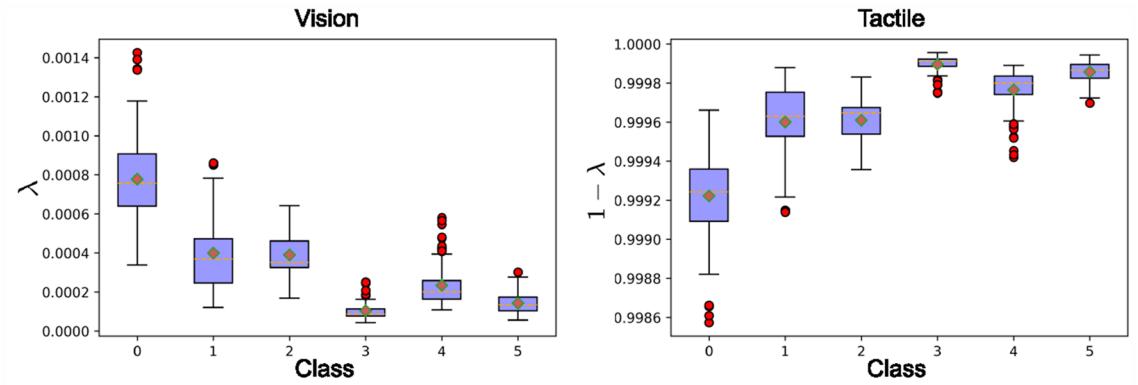
Moreover, the self-adaptive λ-learner can be thought as the fourth kind of attention mechanism used in our model, namely modality attention or branch attention as it makes a trade-off between the output feature vectors of the visual and tactile branches of the network. The classification accuracy gets improvement in all of the three experiments, which is illustrated in Table 7.

5.3 From research to practice

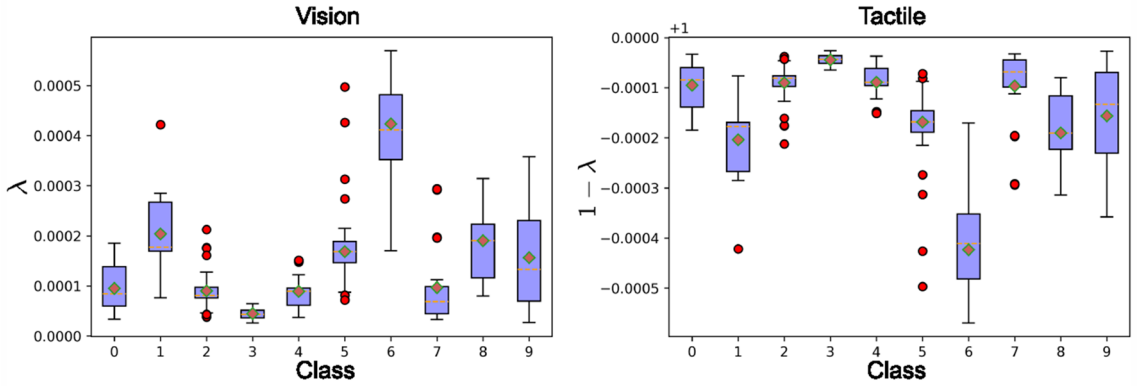
Although our system has shown excellent performance in the experimental evaluation, it has not been promoted to practical application and really brought convenience to the life of visually impaired people. Much more needs to be done to make this continuous, immersive and educational system practical. According to the data from the Department of Educational Planning of China,[1] only about 11,000 people with visual disabilities are in the national education system(or just graduated) in the year of 2020, which is far from the actual number of people in educational age who have vision problems. Many visually impaired people cannot receive compulsory education due to disability, travel restrictions, poverty, etc. Our proposed system has the advantages of low production and maintenance costs, convenient to use, and unsupervised by humans. Therefore, the system has strong promotion potential. When it is deployed in the education place of visually impaired people, with the assorted hardware of the system, it only needs to work with the same shared classification model in the cloud then can achieve high-precision robust classification on a given set of items. The training process for the model's learning new items can be carried out completely on the server side and then deployed to all user terminals, which shows the excellent scalability of the system.

In the next stage, we have plans to apply for funding from the government and public welfare enterprises, to implement the pilot system, and then to promote and apply it to provide solutions for the continuous education of visually impaired people and improve their living ability and life quality. Specifically, we will first join the local government department and information center for the disabled persons to deploy the pilot system to demonstration areas such as the warm homes and schools for the disabled in various districts and counties. Secondly, we will collect the video and tactile data and feedback of visually impaired people using the system, and then based on the collected data and usage feedback, we will further update and improve the system. For example, we might improve the recognition accuracy and generalization ability of the vision and tactile fusion model by introducing the data collected in practical use scene, and design lighter and thinner gloves according to operating habits of visually impaired people. After that, the updated system can be deployed and applied in a new round. Through this deployment—data and feedback collection—system update—redeployment mode, we hope to bring our system from the

[1] http://www.moe.gov.cn.



(a) Experimental results on 6 classes of different bottles.



(b) Experimental results on 10 classes of objects with interference items in view.

Fig. 14 The fusion factors λ and $1-\lambda$ for experiments on dataset(b) and (c). See Table 1 for category labels

Table 7 The improvement in classification accuracy for experiments on different datasets due to the adoption of λ-learner

Dataset	Accuracy/%	
	w/o λ-learner	w/ λ-learner
(a)	99.75	**99.81**
(b)	99.15	**99.42**
(c)	94.77	**95.13**

Bold indicates best performance

research level to practical applications, so as to truly change the lives of visually impaired people.

6 Conclusions

To provide a solution for unsupervised assistance of visually impaired people, we proposed the system AviPer, which aims to be a continuous, immersive, and educational assistance system for visually impaired people to perceive the world. We developed flexible tactile gloves to work with visual recognition to achieve robust multimodal object classification. The key insight of AviPer is that it can provide visually impaired people a sense of participation and real experience in the assistance process, as well as reach a high level of classification accuracy. In the process of developing the system, we overcame the difficulties of heterogeneous data by creatively designing a multi-attention multimodal classification network. We used the intelligent tactile glove to achieve low-cost and stable acquisition of tactile data. We fully respected the privacy of collectors and users in every session from data acquisition to actual application. The verification experiments under various extreme situations prove the robustness of our system. The user experience in the actual scene shows the usability and user-friendliness of the system. We will improve our work in the direction of further improving its generality and putting it into practice. We are looking forward that this work can truly enter the lives of visually impaired people and bring substantial changes to their living ability and life quality.

Acknowledgements This work is supported by the Beijing Nova Program of Science and Technology under Grant Z191100001119129 and the National Natural Science Foundation of China 61702520.

Declarations

Conflict of interest All authors declare no possible conflicts of interest.

References

Ackland, P., Resnikoff, S., Bourne, R.: World blindness and visual impairment: despite many successes, the problem is growing. Community Eye Health **30**(100), 71 (2017)

Ahmetovic, D., Sato, D., Oh, U., Ishihara, T., Kitani, K., Asakawa, C.: Recog: Supporting blind people in recognizing personal objects. In: CHI'20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, pp. 1–12 (2020)

Aladren, A., López-Nicolás, G., Puig, L., Guerrero, J.J.: Navigation assistance for the visually impaired using rgb-d sensor with range expansion. IEEE Syst. J. **10**(3), 922–932 (2014)

Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)

Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: a survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 423–443 (2018)

Banos, O., Villalonga, C., Garcia, R., Saez, A., Damas, M., Holgado-Terriza, J.A., Lee, S., Pomares, H., Rojas, I.: Design, implementation and validation of a novel open framework for agile development of mobile health applications. Biomed. Eng. Online **14**(2), 1–20 (2015)

Barbosa, N.M., Hayes, J., Wang, Y.: Unipass: design and evaluation of a smart device-based password manager for visually impaired users. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, pp. 73–84 (2016)

Björkman, M., Bekiroglu, Y., Högman, V., Kragic, D.: Enhancing visual perception of shape through tactile glances. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3180–3186. IEEE (2013)

Cao, Y., Xu, J., Lin, S., Wei, F., Hu, H.: Gcnet: non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, South Korea (2019)

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. IEEE (2020)

Cui, H., Radosavljevic, V., Chou, F.-C., Lin, T.-H., Nguyen, T., Huang, T.-K., Schneider, J., Djuric, N.: Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 2090–2096. IEEE (2019)

De, D., Bharti, P., Das, S.K., Chellappan, S.: Multimodal wearable sensing for fine-grained activity recognition in healthcare. IEEE Internet Comput. **19**(5), 26–35 (2015)

Delahoz, Y., Labrador, M.A.: A deep-learning-based floor detection system for the visually impaired. In: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), pp. 883–888. IEEE (2017)

Ernst, M.O., Banks, M.S.: Humans integrate visual and haptic information in a statistically optimal fashion. Nature **415**(6870), 429–433 (2002)

Ganz, A., Schafer, J.M., Tao, Y., Wilson, C., Robertson, M.: Percept-ii: Smartphone based indoor navigation system for the blind. In: 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3662–3665. IEEE (2014)

Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., Darrell, T.: Deep learning for tactile understanding from visual and haptic data. In: 2016 IEEE International Conference on Robotics and Automation (ICRA), pp. 536–543. IEEE (2016)

Güler, P., Bekiroglu, Y., Gratal, X., Pauwels, K., Kragic, D.: What's in the container? Classifying object contents from vision and touch. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3961–3968. IEEE (2014)

Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, pp. 7132–7141 (2018)

Hu, D.: An introductory survey on attention mechanisms in nlp problems. In: Proceedings of SAI Intelligent Systems Conference, pp. 432–448 (2019). Springer

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, Ohio, USA, pp. 1725–1732 (2014)

Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

Kroemer, O., Lampert, C.H., Peters, J.: Learning dynamic tactile sensing with robust vision-based training. IEEE Trans. Robot. **27**(3), 545–557 (2011)

Kumar, A., Reddy, S.S.S., Kulkarni, V.: An object detection technique for blind people in real-time using deep neural network. In: 2019 Fifth International Conference on Image Information Processing (ICIIP), pp. 292–297. IEEE (2019)

Lakde, C.K., Prasad, P.S.: Navigation system for visually impaired people. In: 2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), pp. 0093–0098. IEEE (2015)

Lee, J., Jung, Y., Kim, H.: Dual attention in time and frequency domain for voice activity detection. arXiv preprint arXiv:2003.12266 (2020)

Li, F., Neverova, N., Wolf, C., Taylor, G.: Modout: Learning multi-modal architectures by stochastic regularization. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, pp. 422–429 (2017). https://doi.org/10.1109/FG.2017.59

Li, Y., Zhu, J.-Y., Tedrake, R., Torralba, A.: Connecting touch and vision via cross-modal prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, pp. 10609–10618 (2019)

Liang, Y., Ke, S., Zhang, J., Yi, X., Zheng, Y.: Geoman: Multi-level attention networks for geo-sensory time series prediction. In: Twenty-seventh international joint conference on artificial intelligence (IJCAI-18), Stockholm, Sweden, pp. 3428–3434 (2018)

Lin, B.-S., Lee, C.-C., Chiang, P.-Y.: Simple smartphone-based guiding system for visually impaired people. Sensors **17**(6), 1371 (2017)

Liu, H., Guo, D., Zhang, X., Zhu, W., Fang, B., Sun, F.: Toward image-to-tactile cross-modal perception for visually impaired people. IEEE Trans. Autom. Sci. Eng. **18**, 521–529 (2021)

Luo, S., Mou, W., Althoefer, K., Liu, H.: Localizing the object contact through matching tactile features with visual map. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 3903–3908. IEEE (2015)

Metatla, O., Bardot, S., Cullen, C., Serrano, M., Jouffrais, C.: Robots for inclusive play: co-designing an educational game with visually impaired and sighted children. In: CHI'20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, pp. 1–13 (2020)

Morrison, D., Corke, P., Leitner, J.: Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach. arXiv preprint arXiv:1804.05172 (2018)

Neto, L.B., Grijalva, F., Maike, V.R.M.L., Martini, L.C., Florencio, D., Baranauskas, M.C.C., Rocha, A., Goldenstein, S.: A kinect-based wearable face recognition system to aid visually impaired users. IEEE Trans. Hum. Mach. Syst. **47**(1), 52–64 (2016)

Papadopoulos, K.S., Goudiras, D.B.: Accessibility assistance for visually-impaired people in digital texts. Br. J. Vis. Impair. **23**(2), 75–83 (2005)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: an imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)

Poggi, M., Mattoccia, S.: A wearable mobility aid for the visually impaired based on embedded 3d vision and deep learning. In: 2016 IEEE Symposium on Computers and Communication (ISCC), pp. 208–213. IEEE (2016)

Praveen, R.G., Paily, R.P.: Blind navigation assistance for visually impaired based on local depth hypothesis from a single image. Procedia Eng. **64**, 351–360 (2013)

Qin, Y., Song, D., Chen, H., Cheng, W., Jiang, G., Cottrell, G.: A dual-stage attention-based recurrent neural network for time series prediction. arXiv preprint arXiv:1704.02971 (2017)

Rahman, M.W., Tashfia, S.S., Islam, R., Hasan, M.M., Sultan, S.I., Mia, S., Rahman, M.M.: The architectural design of smart blind assistant using iot with deep learning paradigm. Internet Things **13**, 100344 (2021)

Ramachandram, D., Lisicki, M., Shields, T.J., Amer, M.R., Taylor, G.W.: Structure optimization for deep multimodal fusion networks using graph-induced kernels. CoRR arXiv:1707.00750 (2017)

Ramachandram, D., Taylor, G.W.: Deep multimodal learning: a survey on recent advances and trends. IEEE Signal Process. Mag. **34**(6), 96–108 (2017)

Romano, J.M., Hsiao, K., Niemeyer, G., Chitta, S., Kuchenbecker, K.J.: Human-inspired robotic grasp control with tactile sensing. IEEE Trans. Robot. **27**(6), 1067–1079 (2011)

Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)

Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199 (2014)

Smith, J.R., Fishkin, K.P., Jiang, B., Mamishev, A., Philipose, M., Rea, A.D., Roy, S., Sundara-Rajan, K.: Rfid-based techniques for human-activity detection. Commun. ACM **48**(9), 39–44 (2005)

Snoek, C.G., Worring, M.: Multimodal video indexing: a review of the state-of-the-art. Multimed. Tools Appl. **25**(1), 5–35 (2005)

Sun, J., Jiang, J., Liu, Y.: An introductory survey on attention mechanisms in computer vision problems. In: 2020 6th International Conference on Big Data and Information Analytics (BigDIA), pp. 295–300. IEEE (2020)

Sundaram, S., Kellnhofer, P., Li, Y., Zhu, J.-Y., Torralba, A., Matusik, W.: Learning the signatures of the human grasp using a scalable tactile glove. Nature **569**(7758), 698–702 (2019)

Tapu, R., Mocanu, B., Zaharia, T.: Deep-see: joint object detection, tracking and recognition with application to visually impaired navigational assistance. Sensors **17**(11), 2473 (2017)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, pp. 7132–7141 (2018)

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)

Wang, H.-C., Katzschmann, R.K., Teng, S., Araki, B., Giarré, L., Rus, D.: Enabling independent navigation for visually impaired people through a wearable vision-based feedback system. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 6533–6540. IEEE (2017)

Warren, D.H.: Blindness and Children: An Individual Differences Approach. Cambridge University Press, Cambridge (1994)

Wen, Q., Sun, L., Song, X., Gao, J., Wang, X., Xu, H.: Time series data augmentation for deep learning: a survey. arXiv preprint arXiv:2002.12478 (2020)

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, pp. 3–19 (2018)

World Health Organization, et al.: World report on vision (2019)

Wu, D., Pigou, L., Kindermans, P.-J., Le, N.D.-H., Shao, L., Dambre, J., Odobez, J.-M.: Deep dynamic neural networks for multimodal gesture segmentation and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **38**(8), 1583–1597 (2016)

Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., López, A.M.: Multimodal end-to-end autonomous driving. IEEE Trans. Intell. Transp. Syst. 23(1), 537–547 (2020). https://doi.org/10.1109/TITS.2020.3013234

Yin, J., Yang, Q., Pan, J.J.: Sensor-based abnormal human-activity detection. IEEE Trans. Knowl. Data Eng. **20**(8), 1082–1090 (2008)

Yuan, W., Li, R., Srinivasan, M.A., Adelson, E.H.: Measurement of shear and slip with a gelsight tactile sensor. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 304–311. IEEE (2015)

Yuhas, B.P., Goldstein, M.H., Sejnowski, T.J.: Integration of acoustic and visual speech signals using neural networks. IEEE Commun. Mag. **27**(11), 65–71 (1989)

Zangaladze, A., Epstein, C.M., Grafton, S.T., Sathian, K.: Involvement of visual cortex in tactile discrimination of orientation. Nature **401**(6753), 587–590 (1999)

Zhao, Y., Szpiro, S., Knighten, J., Azenkot, S.: Cuesee: exploring visual cues for people with low vision to facilitate a visual search task. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Heidelberg, Germany, pp. 73–84 (2016)

Xinrong Li is an undergraduate student in Tsien Excellence in Engineering Program at School of Aerospace Engineering, Tsinghua University, Beijing, China since 2019. He is currently a research intern in the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China.

His research interests include deep learning, multimodal fusion, and intersection of AI and mechanics.

Meiyu Huang received a B.S. degree in computer science and technology from Huazhong University of Science and Technology, Wuhan, China, in 2010, and a Ph.D. degree in computer application technology from the University of Chinese Academy of Sciences, Beijing, China, in 2016. She is currently an assistant researcher in the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China.

Her research interests include machine learning, ubiquitous computing, human-computer interaction, computer vision and image processing.

Yao Xu received a B.S. degree in electrical and computer engineering in Shanghai Jiao Tong University, China, in 2013 and a M.S. degree in electrical and computer engineering in University of California, Irvine, US, in 2016. He is currently an assistant researcher in the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China.

His research interests include deep learning, data Fusion, distributed system and computer architecture.

Yingze Cao is currently a senior engineer in Qianxuesen Lab, China Academy of Space Technology. She received her M.S. degree(2011) and Ph.D. degree (2016) in Chemistry from Tsinghua University. She joined QianLab in 2016. Her research interests are focused on interfacial materials and smart functional materials.

Yamei Lu is currently a materials engineer in Institute of Flexible Electronics Technology of THU. Zhejiang. She received her M.S. degree (2020) in Chemistry from Beijing JiaoTong University. She used to study and work as a research assistant in Qianxuesen Lab, China Academy of Space Technology (2019-2021). Her research interests include interfacial materials and flexible electronic packaging materials.

Pengfei Wang is currently a Professor in Qianxuesen Lab, China Academy of Space Technology. He received his M.S. degree (2008) and Ph.D. degree (2014) in Mechanics from Xi'an Jiaotong University. His research interests are intelligent structure materials and 3D/4D manufacture.

Xuehsuang Xiang received a B.S. degree in computational mathematics from Wuhan University, Wuhan, China, in 2009, and a Ph.D. degree in computational mathematics from the Academy of Mathematics and Systems Science of Chinese Academy of Sciences, Beijing, China, in 2014. In 2016, he was a postdoctoral researcher with the Department of Mathematics, National University of Singapore, Singapore. He is currently an associate researcher in the Qian Xuesen Laboratory of Space Technology, China Academy of Space Technology, Beijing, China.

His research interests include numerical methods for partial differential equations, image processing and deep learning.