MobileAid: A Fast and Effective Cognitive Aid System on Mobile Devices

XINYI LIU1, BAOYING LIU1,2, GUOQING LIU1, FENG CHEN1, AND TIANZHANG XING1,2, (Member, IEEE)

1School of Information Science and Technology, Northwest University, Xi’an 710127, China
2Shaanxi International Joint Research Centre for the Battery-Free Internet of Things, Northwest University, Xi’an 710127, China

Corresponding author: Tianzhang Xing (xtz@nwu.edu.cn)

This work was supported in part by the Key Research and Development Project of Shaanxi Province under Grant 2018SF-369, in part by the China Postdoctoral Science Foundation under Grant 2017M613187, in part by the International Cooperation Project of Shaanxi Province under Grant 2020KW-004, and in part by the Shaanxi Science and Technology Innovation Team Support Project under Grant 2018TD-026.

ABSTRACT Elderly people generally suffer from diseases with cognitive decline, which seriously affect their daily lives. To assist these patients with cognitive decline, numerous cognitive aid devices have been researched and designed. Although these devices can achieve the aid tasks, they have to face two difficulties: low accuracy and long latency. In this paper, we present MobileAid, an aid system implemented on the mobile device, which assists cognitive decline patients to recognize objects. The key idea of this system is a two-step lightweight neural network design for the target recognition: context recognition and object recognition, and this design achieves low time delay. Considering the great success of convolutional neural networks in object recognition, we design a lightweight convolutional neural network by the combination of pooling layers and activation functions selection, which achieves high accuracy. Furthermore, we apply the depthwise separable convolution to reduce the resource consumption for deploying the system on mobile devices. The results of the extensive experiments we have conducted show that MobileAid can achieve 95% high accuracy and 90ms low time delay with few resource consumption.

INDEX TERMS Deep learning, mobile computing, context recognition, object recognition.

I. INTRODUCTION

As is known to all that the disease of cognitive decline is prevalent among elderly people. They often suffer from memory lapses or even loss, at the same time, they are unable to recognize their surroundings and objects [1]. According to the statistics of the world population, the number of people over 60 years of age is increasing in many parts of the world, and about a quarter of them suffers from cognitive decline. In addition, patients with cognitive decline may be suffering from Alzheimer, change of behavior, and degeneration of brain function [2], [3]. This disease has already attracted the attention of researchers, governments and the society [4].

In order to assist patients with cognitive decline, tremendous efforts have been made [5]. Although these patients are unable to recognize many scenes and objects that they were familiar with, they can still understand simple basic input such as simple words or sounds. Therefore, objects, people or scenes that are elusive for the patients are presented directly to them in a simple and recognizable form. Based on these characteristics of the patients, previous works have focused on implementing special sensors in areas where the patient is frequently active. However, these methods have big disadvantages, such as the narrow range of sensors and the large cost of this arrangement [6], [7]. In addition, there are many other tasks in medical care. For instance, guarding sleep via WiFi or RF signals, detecting arrhythmias, and monitoring breath [38]–[42]. There are also methods of monitoring the posture of the human gesture to observe the status of patients [44], [45].

With the rise of wearable devices and their increased use in recent years, researchers are shifting their focus from sensors to wearable devices, such as Gabriel that get a great success [8]. By deploying the cognitive assistance system on the wearable device to capture the patient’s first-person perspective, and then unloading the video stream obtained from this perspective on the cloud or server for further processing, the object recognition of the first-person perspective is finally
completed [35]. Eventually, the recognition results will be translated into words or sounds that patients can understand at ease, and then apply these forms to the wearable devices [43].

The advent of 5G technology has also facilitated the development of this approach, with data transfer time between local and server being negligible. However, the cost of reducing the footprint of local processing resources is the loss of user privacy during data transfer [9], [10]. Because in the process of data offloading, information is vulnerable to attack, resulting in incomplete data. Offloading is a mechanism that allows us to close the gap by performing intensive calculations on a large system with sufficient resources for the application [52]–[56]. An emerging paradigm to address this is the deployment of systems to mobile devices, which greatly improves the experience of privacy-sensitive users. From servers to mobile devices, the demand for computing and energy is growing. Because mobile devices are often limited in terms of computing power, storage capacity and energy consumption [46]. And this technology is also widely used in scenarios other than medical care, such as smart cities, smart transportation and smart homes [47]–[51].

Although these designs can effectively assist patients to recognize targets, with the advancement of technology and the growing of patients’ needs, existing cognitive assistance systems still have the following limitations.

1) The system should be able to recognize the targets in near real time. In order to meet the needs of patients to obtain recognition results from a short interval after seeing objects, the system delay is expected to be within 300ms [11]. It is well known that video streams cannot be processed well on mobile platforms with limited computing resources. Therefore, it is better to offload the video stream to computationally powerful devices like the cloud or the server, but this will inevitably add to the latency and fail to meet the patient's recognition needs.

2) The system must propose the real accurate recognition results. There are two main ways to accomplish this task: computer vision and deep learning. For instance, the low accuracy is a problem in the computer vision method, and the SURFTrac is the most common among the computer vision methods [12], but the result of recognition is less than 50%. When complex networks are used, such as VGG16, AlexNet, InceptionNets [13]–[15], these neural networks achieve high accuracy due to their deeper neural network architectures. However, the complexity of neural network structure will lead to the burden of local resources and the increase of computation time.

According to the sense of FIGURE 1, patients are equipped with wearable devices in their familiar environment. While viewing the surroundings from a first-person perspective, the camera of the device receives a stream of first-person video from the patient, and then the device processes to communicate the identified object or person to the patient in a form that the patient can easily understand. In order to address the limitation of time delay and accuracy, we propose a mobile aid system for the cognitive decline patients in this paper, named MobileAid. For the limitation of time delay, we deploy the network of system on the mobile platform and process the data locally, which not only eliminates the need for frequent data exchange, but also protects the user's first view data. The next limitation is low accuracy. As is known to all, convolutional neural network (CNN) performs well in object recognition, so we design our recognition methods based on this characteristic. Our ultimate goal is to improve the accuracy of object recognition by using the neural network design, and we carefully study the structure of each part of the neural network, including the functions and characteristics of each layer. Finally, our research and experimental results show that choosing the combination of different pooling layers and the collocation of different activation functions can effectively improve the final accuracy.

However, we face a major challenge in our system design. Since the convolutional neural network is computation-intensive and resource-intensive, the computing power of mobile platforms, the memory they can occupy and the energy they can consume are limited. For instance, a commonly used neural network, VGG16, has thirteen convolutional layers and takes about 100 seconds to process a single image using CPU on Samsung Galaxy S7 [9]. Therefore, what really caught our attention is how to deploy neural networks on mobile platforms. We must reduce the amount of computation and network size of the neural network so as to achieve the goal of reducing energy consumption. To address this challenge, we have made a series of efforts to optimize the neural network.

The general recognition system uses a generic neural network for object recognition, which undoubtedly increases the search space and network scale. It also adds to the difficulty...
of deploying on resource-constrained platforms. We consider that the environment in which each patient lives is usually limited, so we first judge the contextual information from the first perspective of the patient, and then identify the object. In this way, it greatly reduces the search space and neural network scale of recognition. So we split the large network into a lightweight context neural network and an object neural network. First, we get the context of the patient, reducing the number of objects that can be matched, and then we use the object network to realize the final recognition.

In the convolutional neural network, the part that consumes the most computation is the convolutional layer, so we choose to optimize the convolutional layer. We find a convolutional method that reduces the amount of computation and the number of parameters, named depthwise separable convolution, which makes the convolution into two steps, deep-wise convolution and point-wise convolution. We have replaced the standard convolution of the network in this way.

In conclusion, this paper has the following contributions:

- We propose a two-step lightweight neural network to treat the object recognition task as a context recognition task and an object recognition task. This design perfectly addresses the challenge that deep learning networks cannot be deployed well on resource-limited platforms. This idea can also be generalized to other models.
- Based on achieving lower time delay, we design a lightweight neural network by using a set of optimization methods, such as the combination of pooling layers, the selection of activation functions and the depthwise separable convolution. These details can make an excellent trade off between accuracy and resource.
- We develop the MobileAid and conducted extensive experiments with public datasets and our own dataset. It finally achieves 95% high accuracy and 90ms low time delay with few resource.

In the following chapters of this paper, we will introduce the background, the overview of our system, the detail of MobileAid design, the experiments results to proof the effectiveness of MobileAid, and the conclusion and future expectation.

**II. BACKGROUND**

**A. RECOGNITION METHODS COMPARISON**

SURF and CNN are the most commonly used to recognize images. SURF algorithm is the accelerated version of SIFT algorithm. It is mainly used to match objects in two images to complete image recognition. Based on this feature of SURF algorithm, it is often used for image recognition and 3D reconstruction. But the low accuracy of the recognition is a significant disadvantage of this approach, as SURFTrac can achieve less than 50% accuracy. So related works think of using deep learning to recognize images, such as AlexNet achieves 80.8% top-5 accuracy, VGG16 achieves 90.5% top-5 accuracy, and ResNet achieves 94.3% [13], [14], [16]. Although these have greatly improved the accuracy, the model level is deeper and the computing or memory requirements are bigger. These very deep models cannot be applied in real time on mobile platforms.

**B. THE EFFECTIVENESS OF DNNS ON MOBILE PLATFORM**

In recent years, a number of researches have shown the desirability of deploying DNN on mobile platforms, such as DeepEar, DeepX, and DeepMon. The main work of DeepEar is to implement the audio sensor on mobile DSP, which is a power-limited platform [17]. This work demonstrates the feasibility that sensors can be placed on resource-limited platforms. The main contribution of DeepX is to reduce the hardware resources required for deep learning inference and enable it to be deployed on mobile platforms [18]. It breaks up the whole neural network into blocks and hands them over to different local processors. However, these require powerful hardware. In addition, DeepMon implements low-latency deep DNN on mobile GPU through various optimization techniques [9]. But its combination of delay and precision can only be achieved through expensive cloud servers with very good network connections. These needs cannot be met in a resource-constrained cognitive aid system.

**C. REDUCTION OF THE NETWORK MODEL**

Existing works generally speed up inference by reducing the scale of network models, which are called light-weight networks. For example, pruning, weight quantization, Huffman coding, volume integration solutions, and special layer structures proposed by Han et al. [19]. And there are various convolution optimization methods, like depthwise separable convolution [20], [21], group convolution [22] and dilated convolution [23]. And we choose from the active layer, pool layer and convolution layer to propose a series of optimized combination methods.

**III. SYSTEM OVERVIEW**

FIGURE 2 shows the framework and signal processing flow of MobileAid system. Through this framework, the system principles and objectives of our requirements can be met. The framework of MobileAid mainly consists of two parts, one is context recognition, and the other is object recognition. According to the characteristics of context and object recognition, we design two different lightweight neural networks. The mobile device first obtains the first person video stream, and then begins a series of data pre-processing, such as image select. This part will select the video frame containing valid information from the first person video stream. Then the selected video frame is sent to the context network. The design of context network includes the combination of pooling layers and the selection of activation functions. The context network searches the context database for the scene where the focused object is. If the search is successful, the object will be recognized by using the object neural network, which is trained by the current focused scene. And the search space of the neural network is no longer filled with thousands of objects, but the objects that may appear in the
current scene, and the number of searching objects is greatly reduced.

**A. CONTEXT RECOGNITION NETWORK**

The first module to be executed is the context recognition network. The premise of our context recognition module is to process the video stream. Compared with many common methods to extract video frames, we finally use the accelerometer to select video frames within a certain interval. Then the image is handed over to the context network for processing. The main function of this network is to reduce the search space of the object recognition network so as to maximize the effect of the lightweight neural network. This means that the number of objects recognized by the object network is determined by the output of the context network.

As is known to all, existing recognition methods focus on computer vision technologies such as SURF and the complex deep neural networks such as VGG16 and AlexNet [24], [25]. According to the experience, the accuracy and delay performance of recognition of computer vision technology are far less than that of deep learning. However, using complex neural networks requires resources and energy consumption that mobile devices cannot afford. Inspired by the above researches, we design a lightweight neural network based on the daily living environment of patients with cognitive decline to meet the requirements for deployment on mobile devices.

To verify that we have improved the effectiveness of the context-aware network, we use two datasets to test the network with and without context information, in which the Measure dataset was our own dataset. As can be seen from FIGURE 3, the network with context information can improve the recognition accuracy by 15%-30% compared with the network without context information. It is further observed that the recognition accuracy decreases with the increase of object types. Because in these two datasets, there are a large number of objects that do not contain context information, which reduces the accuracy of context recognition network. But this does not affect our proof of the validity of context network.

**B. OBJECT RECOGNITION NETWORK**

In the former stage, through the context recognition network, we have obtained the context information of the patient, which reduces the search space of object recognition. Helping the patient to accurately identify the object in focus is the ultimate goal of MobileAid. From SURF to complex neural networks like InceptionNets and others, there are many aspects we need to consider. Among them, the accuracy obtained by using the neural network with high complexity is very high, generally higher than 80%. The price of high accuracy is increased the computing costs and platform consumption of resources, which undoubtedly limit the deployment of complex networks on resource-limited platform. Therefore, in order to successfully deploy the network on the mobile platform and achieve high accurate, we use the convolution optimization method based on the context lightweight neural network. It not only keeps the accuracy of recognition, but also reduces latency, memory and power consumption.

**IV. SYSTEM DESIGN**

**A. IMAGE SELECT**

Our system is implemented on an energy-constrained platform that explores the potential for optimizing the use of limited computing resources and energy through in-depth study of selected images. First, in our system, we assume that the patient is browsing the world using a mobile or wearable device whose camera generates a first-person video stream. However, we can not process the entire video stream, but a few video frames that contain valid information because of the resource-limited mobile platform. Therefore, our work is to design an efficient way to select video frames.
There are many ways to select an image from a video stream, of which two are the most common. The first view is that the first-person perspective changes little or nothing in a short period of time [9]. The experiments of this work show that the interval between two image frames is basically unchanged at 500ms. Another method is to use the accelerometer to determine when to select an image frame [26]. Because acceleration can demonstrate whether the first-person perspective has stopped or not. When the perspective of the patient changes, the patient may be browsing the context, so he does not need to get accurate information about the object. Conversely, if the user is focusing on the object he wants to identify, the accelerometer measures zero. Therefore, this video frame contains the valid information we need for the study. So we can choose this video frame.

We deploy the system on a mobile device with limited resources, but both methods consume a lot of computer resources and cannot be adopted. The first method to select frames at 500ms cannot determine whether the selected video frame contains valid information, or whether the selected video frame may be the vaguest image in 500ms. And the second approach has the potential to produce a large number of redundant frames because it ignores the similarity between successive video frames and wastes the limited resources of the mobile device.

These existing methods have their own obvious advantages and disadvantages, and we finally decide to combine the advantages of them. We first determine the time interval between images in the video stream, which is 500ms. Then we use the accelerometer measurements to determine which frame in the 500ms to select. This method can effectively overcome the problem of video frame processing on mobile platform.

### B. NETWORK ARCHITECTURE

Using convolutional neural network to recognize the image is a deep learning method, which is widely used in video, image and speech recognition [27], [28]. Due to the emergence of convolutional neural network, the image recognition ushered in the peak. The network consists of different layers, each has its own functions and characteristics. The function of the convolutional layer is to extract the features of image, which is the most important layer. The pooling layer reduces the redundant information of the previous layer, thus reducing the amount of computation. The activation layer eliminates the linear expression of the previous layer’s output through the activation function, which enriches the network’s output. The full connection layer usually classifies the images with category probability at the end of a network. Therefore, our basic network is designed in combination with the principle and target of the CNN itself.

We design a two-step lightweight neural network, one for contextual information and one for object recognition. Since the purpose of the two networks is different, the architecture of the network is also different, which is shown in TABLE 1.

### TABLE 1. The architecture of the two-step network.

| Layer | Object Network | Context Network |
|-------|----------------|-----------------|
|       | Size Out | Kernel | Size Out | Kernel |
| conv1 | 52*32*64 | 5*5,1 | 52*32*128 | 5*5,1 |
| pool1 | 15*15*64 | 3*3,2 | 15*15*128 | 3*3,2 |
| conv2 | 15*15*64 | 5*5,1 | 15*15*64 | 5*5,1 |
| pool2 | 7*7*64 | 3*3,2 | 7*7*64 | 3*3,2 |
| conv3 | 7*7*128 | 5*5,1 | 7*7*64 | 5*5,1 |
| pool3 | 2*2*128 | 5*5,1 | 2*2*64 | 5*5,1 |
| fc1   | 64 | - | 64 | - |
| fc2   | 32 | 32 | - |
| fc3   | 10 | 5 | - |

1) **THE COMBINATION OF POOLING LAYERS**

Convolution operation is often associated with a lot of redundant information. For this reason, a pooling layer is designed after the convolutional layer. And its most important function is to reduce the redundancy brought by this kind of operation, so as to reduce the output error of the convolutional layer. There are two types of errors. First of all, the size of the pooling region of each pooling layer is uncertain, and this value will affect the size of the difference. Secondly, the parameter setting of convolutional layer will also produce errors to the result. We chose average pooling (AVE) and maximum pooling (MAX) to address these two problems, respectively. As a matter of course, different methods of pooling layers produce different results.

Finally, we test the performance of eight different combinations. The test results show that the combination of three average pooling has the best performance, whose accuracy is 90%.

2) **THE SELECTION OF ACTIVATION FUNCTIONS**

We find that the application of the activation function will affect the results of target recognition. Usually, the activation function is added after the neuron to adapt the neuron to more complex nonlinear problems. Sigmoid and ReLU functions are common activation functions in CNNs. The Sigmoid is defined as:

\[
S(x) = \frac{1}{1 + e^{-x}}
\]  

The output of Sigmoid function between (0, 1), the derivation is easy and is physically closest to the biological neuron. However, it also has obvious defects. Because of the soft saturation of the gradient, it is easy to cause the disappearance of the gradient, which leads to the problem of CNN training. Therefore, we choose the ReLU function [29], [30] as the activation layer as follows:

\[
f(x) = \max(0, x)
\]

Compared with Sigmoid function, it can be found that when x is less than 0, it is hard for ReLU to saturate, and when x is greater than 0, there is no problem with saturation. The ReLU function can converge quickly in SGD, effectively alleviating the problem of gradient disappearance. However, simply using the ReLU function for the activation operation,
during the training process, there may be some cases where the neurons die and the weight cannot be updated [31], [32]. Then the gradient through this neuron will always be zero from this point, which means that the neurons will irreversibly die during training.

To address these shortcomings, some ReLU-based functions, such as ELU [33], [34], have been proposed in deep learning.

The ELU function combines the sigmoid function with the ReLU function, and its region less than 0 effectively improves the weakness of ReLU, which does not have strong robustness to the input. The linear part on the right side is not saturated, which can alleviate the disappearance of the gradient. As a result, the output of ELU is close to zero, and the convergence speed is faster.

\[
ELU(x) = \begin{cases} 
  x & \text{if } x \geq 0 \\
  \gamma \cdot (exp(x) - 1) & \text{if } x < 0
\end{cases}
\]

Therefore, we propose a combination of ReLU function and ELU function to realize the complementarity of the two. To verify the feasibility of this theory, we test the combination of the eight activation functions, and the results show that the combination of ReLU + ELU + ELU does get the best performance. TABLE 2 shows the accuracy of different combinations.

**TABLE 2. The combination of activation layer.**

| Combination | Accuracy |
|-------------|----------|
| ReLU+ReLU+ReLU | 93.67% |
| ELU+ELU+ELU | 93.84% |
| ReLU+ELU+ReLU | 89.13% |
| ELU+ELU+ReLU | 91.36% |
| ELU+ReLU+ReLU | 93.70% |
| ReLU+ReLU+ELU | 92.93% |
| ELU+ReLU+ELU | 93.86% |
| ReLU+ReLU+ELU | 95.57% |

C. CONVOLUTION OPTIMIZATION

Although the network designed above achieves high accuracy of this system, the resources of the mobile platform are limited. We need to further consider a suite of optimizations for trading off between energy consumption and accuracy on a resource-constrained platform. It is empirically derived that the computational burden of convolutional layer is the largest in the whole network. Therefore, we consider that optimizing the convolutional layers to achieve accelerating calculations and reducing the amount of parameters. Traditional convolution is the process of extracting features by convolving both regions and channels, and we try to separate the regions from the channels. Inspired by some related works [20], [21], we replace the standard convolution in MobileAid with a channel convolution and a point convolution, called depthwise separable convolution. This convolution greatly reduces the amount of computation and the number of parameters.

The traditional convolution uses kernel and channel as a step to convolve the input image to obtain a feature map as the output. However, the depthwise separable convolution divides a step into two, one for extracting image features and the other for combining the extracted features. And the process of depthwise separable convolution is shown in FIGURE 4. First of all, it captures the features of each channel of the input to get a set of channel feature maps; and then it uses the point convolution to convolute the channel features of the former; and finally, the result is obtained.

We assume that the convolution kernel is \( n^2 \times M \times N \), the input image is \( F_{in}^2 \times M \) and output feature map is \( F_{out}^2 \times N \), where \( M \) and \( N \) are the number of input and output channels, respectively. Then the computational cost of standard convolution operation is

\[
n^2 \times M \times N \times F_{in}^2
\]

If we use depthwise separable convolution, the computational cost is

\[
n^2 \times M \times F_{in}^2 + F_{in}^2 \times M \times N
\]

where the computational cost of the channel convolution is \( n^2 \times M \), and \( F_{in}^2 \times M \times N \) is about the point convolution. And the theoretical computational cost of the standard convolution operation and the depthwise separable convolution are as follows:

\[
n^2 \times M \times F_{in}^2 + F_{in}^2 \times M \times N = \frac{1}{N} + \frac{1}{n^2}
\]

The amount of parameters using the standard convolution is \( n^2 \times M \times N \). And the amount of parameters using the depthwise separable convolution is \( n^2 \times M + M \times N \). Then the parameter amount of the depthwise separable convolution and the standard convolution is:

\[
n^2 \times M + M \times N = \frac{1}{N} + \frac{1}{n^2}
\]

This paper replaces the original standard convolution module with a depthwise separable convolution, which greatly
reduces the computational cost while ensuring that the accuracy decrease is negligible. And the details will be shown in Section V.

V. EXPERIMENT
A. SETUP
1) DATASET
In our experiment, we use three datasets, including CIFAR-10, MIT Place2 and Measure. First, the CIFAR-10 dataset consists of 60,000 32*32 color images and is divided into 10 classes, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck. Each class contains 6,000 images, and there are 50,000 images in the training set and 10,000 images in the test set. Secondly, MIT Place2 is a dataset dedicated to context recognition that contains more than 10 million images, including over 400 unique scene classes. Because the number of scenes in MIT Place2 is so large, there are a lot of places that people do not go to frequently. Finally, we select 5 scenes with the highest probability where people visit called Measure, and each scene has 100 images of 32*32. The five scenes are apartment, office, dining hall, campus, and store. In addition, there are 80 images in training set, and 20 images in test set. The detailed setup for each dataset will be explained later in the experiment.

2) NETWORK STRUCTURE
Our main idea is a two-step network, and each network has different architecture because of their different functions. We get the network architecture after a series of parameter optimization experiments. First is the combination of the pooling layers, next is the selection of activation functions. Finally, due to the high memory and energy consumption requirements of mobile devices, we replace the traditional convolution in the two-step network with the depthwise separable convolution.

We refer to the network that does not use a depthwise separable convolution as Full-MobileAid, and the network with depthwise separable convolution as MobileAid. Taking the object recognition network as an example, the differences between the two networks are shown in TABLE 3. In addition, as mentioned in the table means depthwise separable convolution.

3) PLATFORM STATEMENT
We choose the Lenovo xiaoxin15 as the experimental platform, because the CPU performance and RAM of this laptop are similar to that of LeEco lepro 3. In particular, xiaoxin15 has 8G of RAM, which is the same as a phone. In addition, this laptop uses an i7-8565U dual-core CPU with a frequency of 1.99Ghz. The comparison of the LeEco lepro 3 and this laptop are shown in TABLE 4.

TABLE 4. The specs of LeEco lepro3 and laptop.

| Platform        | CPU     | RAM(G) |
|-----------------|---------|--------|
| Lenovo xiaoxin15| 4       | 1.99   |
| LeEco lepro3    | 4       | 2.35   |

FIGURE 5. Performance of MobileAid on context recognition.

B. CONTEXT RECOGNITION NETWORK
In MobileAid, the first step in our two-step network is the context network. We test the accuracy, time delay, memory and energy consumption of the current network, and then compare the network with other typical deep networks and SURF algorithms. The SURF algorithm mentioned here is not a feature descriptor, but an algorithm proposed in a related work [36]. The SURF algorithm is divided into two categories according to the size of the image, one category is used to recognize the size 128*128, and the other is 256*256. In addition, the network named LeNet-5 comes from another work [37], which is a simplest five-layer network. Since our network is a lightweight network with fewer layers, we take this five-layer network as baseline. We use MIT Place 2 dataset to test the performance of accuracy. As is shown in FIGURE 5, both SURF and LeNet-5 can only achieve 20% accuracy. Because SURF has limitations in image recognition, and LeNet-5 does not have a deep network level, both of
them are perform poorly. Then the accuracy of ILSVRC2015, VGG16 and MobileAid are basically the same, above 80%. MobileAid does not perform the best, and the reason why the other two networks can achieve better performance is that they have deeper level, but the more complex the network will generate more parameters which occupy more platform resources. Considering the resource-constrained demand of mobile devices, MobileAid is more in line with the requirements of our system design. At this point, the matching object of context recognition is all the scenes. Although the accuracy of context recognition is only about 80%, it has nothing to do with the accuracy of object recognition network behind. Because the network used for object recognition is trained for the corresponding scene, the search space is only for the object in the current context.

Then we test the delay, memory and energy consumption of MobileAid, VGG16 and SURF. As we can see in FIGURE 5(b) that the time delay of MobileAid is the lowest among all control groups. Because a different pooling collocation is adopted in the network architecture to reduce the information redundancy after convolution operation. FIGURE 5(c) and FIGURE 5(d) demonstrate that MobileAid has the lowest memory and energy consumption, while memory consumption represents the ratio of memory consumption to total system memory.

C. OBJECT RECOGNITION NETWORK

To verify the effectiveness of the object recognition network, we run our network on the CIFAR-10 image dataset and Measure dataset. In the two-step network, we replace the traditional convolution module with a depthwise separable convolution module. In order to distinguish from MobileAid, we call the network without depthwise separable convolution the Full-MobileAid. First, we use the CIFAR-10 dataset on Full-MobileAid, MobileAid, MobileNet and ImageNet on AlexNet. Secondly, we use the Measure dataset for all these networks. In addition, we have to conduct control experiments with SURF algorithm, so we process the size of some images in the CIFAR-10 dataset accordingly.

| Model         | Accuracy | Million Parameters |
|---------------|----------|--------------------|
| AlexNet       | 85       | 60 M               |
| MobileNet     | 80       | 3.25 M             |
| SURF128       | 35       | -                  |
| SURF256       | 40       | -                  |
| LeNet-5       | 65       | -                  |
| Full-MobileAid| 79       | 2.89 M             |
| MobileAid     | 77       | 0.42 M             |

We first run networks on CIFAR-10, according to TABLE 5, deep neural networks can achieve about 40% accuracy performance better than the SURF. AlexNet gets the best performance of accuracy because of its deeper level, but its million parameters is the largest above all. Although the accuracy of MobileAid is a few percentage points lower, its parameter performs much better than AlexNet, which can completely compensate for the lack of accuracy in MobileAid. Meanwhile, because of the addition of depthwise separable convolution, we compare the performance of Full-MobileAid, MobileAid and MobileNet, the accuracy of three is basically the same. Although MobileAid performs slightly worse, the decrease in accuracy is negligible compared to the large reduction in the amount of parameters.

![FIGURE 6. The accuracy comparison of MobileAid on object recognition.](image)

FIGURE 6(a) shows the accuracy results of each network and algorithm when using the Measure dataset. Since this dataset was carefully selected, the networks show better performance. We also perform the accuracy of MobileAid in five scenarios in the Measure dataset, and the results are shown in the FIGURE 6(b). Compared with store, dining hall and campus, apartment and office have a lower accuracy. Because store, dining hall and campus are public places, and objects in different scenes have little difference. As apartment and office are personalized places, the objects in the scene are different, and the types of objects are more complex, so it will be more difficult to identify and the accuracy will be relatively low.

![FIGURE 7. Performance of MobileAid on object recognition.](image)

FIGURE 7(a) shows the result of time delay. We find that the longest time delay is AlexNet, nearly 400ms, and the SURF algorithm only reaches about two-thirds of AlexNet. On the contrary, the MobileAid system performs best. And the result of memory consumption is shown in FIGURE 7(b), AlexNet also stands out from other methods. These demonstrate that the complex network will cost more resources and energy, and these costs are unaffordable on mobile platforms. So even if the accuracy of AlexNet is the highest,
we do not consider these complicated networks. Concerning LeNet-5 and MobileAid, although LeNet-5 has the least delay, MobileAid’s memory consumption and accuracy are much higher than LeNet-5. As for SURF128 and SURF256, the performance is general in terms of delay and memory. Therefore, MobileAid can perform better when we implement it on a mobile platform with limited resources.

Finally, the MobileAid system is implemented on the mobile devices. We deploy the trained networks on the mobile phone, as shown in FIGURE 8. We can recognize the mouse, which usually appears in the office context.

FIGURE 8. MobileAid implemented on mobile device.

VI. CONCLUSION AND FUTURE WORK

We propose a fast and effective aid system for cognitive decline elderly people called MobileAid. Considering the influence of search space on network complexity, we design a two-step lightweight neural network with context information. By analyzing the architecture of the network and the problems existing in the current cognitive assistance system, we design the context network and the object network respectively by using different combinations of pooling layer and activation layer, and the depthwise separable convolution. Experiments show that our system can be implemented on a mobile platform with limited resources, which can achieve object recognition in almost real time, with high accuracy and low resource consumption. Ultimately, MobileAid achieves the desired goal.

We propose MobileAid, which is a prototype system, and there is a lot we can do to improve our system. Firstly, we will strengthen the data match of database. The database stores the context in the database that cognitive decline patients often arrive. And the network we designed will match the object context with the context in the database. If the match is successful, the data of the context will read in order to wait for the next step. Otherwise, the database will be updated. Secondly, we will optimize our network structure to further reduce latency and energy consumption. Thirdly, the accurate recognition of multi-objective recognition is our main work. Finally, the main design idea of our two-step network is to reduce the search space by decomposing tasks. This allows us to achieve high recognition rates even with lightweight networks. This idea is not limited to helping people with cognitive decline recognize objects, but it could also be applied to other areas where mobile devices are needed to recognize objects with contextual information.

REFERENCES

[1] S. Hodges, E. Berry, and K. Wood, “SenseCam: A wearable camera that stimulates and rehabilitates autobiographical memory,” Memory, vol. 19, no. 7, pp. 685–696, Oct. 2011.
[2] J. F. Sumowski, M. A. Rocca, V. M. Leavitt, J. Dackovic, S. Mesaros, J. Drulovic, J. DeLuca, and M. Filippi, “Brain reserve and cognitive reserve protect against cognitive decline over 4.5 years in MS,” Neurology, vol. 82, no. 20, pp. 1776–1783, May 2014.
[3] G. W. Small, L. M. Erroli, D. H. Silverman, S.-C. Huang, S. Komo, S. Y. Bookheimer, H. Lavretsky, K. Miller, P. Siddarth, N. L. Raagun, J.-C. Marziotti, S. Saxena, H. M. Wai, M. S. Mega, J. L. Cummings, A. M. Saunders, M. A. Pericak-Vance, A. D. Roses, J. R. Barrio, and M. E. Phelps, “Cerebral metabolic and cognitive decline in persons at genetic risk for Alzheimer’s disease,” Proc. Nat. Acad. Sci. USA, vol. 97, no. 11, pp. 6037–6042, 2000.
[4] G. G. Glenner and C. W. Wong, “Alzheimer’s disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein,” Biochem. Biophys. Res. Commun., vol. 120, no. 3, pp. 885–890, May 1984.
[5] P. C. Roy, S. Giroux, B. Bouchard, A. Bouzouane, C. Phua, A. Tolstikov, and J. Biswas, “A possibilistic approach for activity recognition in smart homes for cognitive assistance to Alzheimer’s patients,” in Activity Recognition in Pervasive Intelligent Environments, Paris, France: Atlantis Press, 2011.
[6] S. Helal, W. Mann, H. El-Zabadani, J. King, Y. Kaddoura, and E. Jansen, “Gator tech smart house: A programmable pervasive space,” Computer, vol. 38, no. 3, pp. 50–60, Apr. 2005.
[7] V. Stanford, “Using pervasive computing to deliver elder care,” IEEE Per. Comput., vol. 1, no. 1, pp. 10–13, Jan. 2002.
[8] K. Ha, Z. Chen, W. Hu, W. Richter, P. Pillai, and M. Satyanarayanan, “Towards wearable cognitive assistance,” in Proc. 12th Annu. Int. Conf. Mobile Syst., Appl., Services MobiSys, 2014, pp. 68–81.
[9] L. N. Huynh, Y. Lee, and K. R. Balan, “DeepMon: Mobile GPU-based deep learning framework for continuous vision applications,” in Proc. 15th Annu. Int. Conf. Mobile Syst., Appl., Services, Jun. 2017, pp. 82–95.
[10] A. Aleem and C. Ryan Sprott, “Let me in the cloud: Analysis of the benefit and risk assessment of cloud platform,” J. Financial Crime, vol. 20, no. 1, pp. 6–24, Dec. 2012.
[11] T. R. Agus, S. J. Thorpe, C. Suied, and D. Pressnitzer, “Characteristics of human voice processing,” in Proc. IEEE Int. Symp. Circuits Syst., May 2010, pp. 509–512.
[12] D.-N. Ta, W.-C. Chen, N. Gelfand, and K. Pulli, “SURFTrac: Efficient tracking and continuous object recognition using local feature descriptors,” in Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR), Jun. 2009, pp. 2937–2944.
[13] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556. [Online]. Available: https://arxiv.org/abs/1409.1556
[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2012, pp. 1097–1105.
[15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 1–9.
[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.
[17] N. D. Lane, P. Georgiev, and L. Q. Dendro, “DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning,” in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Ubicomp, 2015, pp. 285–294.
[18] N. D. Lane, S. Bhattacharya, P. Georgiev, C. Forlivesi, L. Jiao, L. Q. Dendro, and F. Kawser, “DeepX: A software accelerator for low-power deep learning inference on mobile devices,” in Proc. 15th ACM/IEEE Int. Conf. Int. Process. Sensor Netw. (IPSN), Apr. 2016, pp. 1–12.
[19] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding,” in Proc. 4th Int. Conf. Learn. Represent., ICLR, San Juan, Puerto Rico, May 2016, pp. 1–14.
[20] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1251–1258.
[21] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, “Mobilenets: Efficient convolutional neural networks for mobile vision applications,” 2017, arXiv:1704.04861. [Online]. Available: https://arxiv.org/abs/1704.04861
[22] G. Huang, S. Liu, L. V. D. Maaten, and K. Q. Weinberger, “Con- denseNet: An efficient Densenet using learned group convolutions,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2752–2761.
[23] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, “ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmen- tation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 558–562.
[24] S. Chen, C.-D. Wu, X.-S. Yu, and D.-Y. Chen, “Fast scene recognition based on saliency region and SURF,” in Proc. 2nd Int. Conf. Intell. Control Inf. Process., Jul. 2011, pp. 863–866.
[25] L. Wang, G. Sheng, W. Huang, and Q. Yu, “Places205-VGGNet models for scene recognition,” 2015, arXiv:1508.01667. [Online]. Available: https://arxiv.org/abs/1508.01667
[26] M. Rahman, U. Topkara, and B. Carbunar, “Moveec: Video liveness veriﬁcation for mobile devices using built-in motion sensors,” IEEE Trans. Mobile Comput., vol. 15, no. 5, pp. 1197–1210, May 2016.
[27] T. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” Biol. Cybern., vol. 36, no. 4, pp. 193–202, Apr. 1980.
[28] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Neural Comput., vol. 1, no. 4, pp. 541–551, Dec. 1989.
[29] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, May 2015.
[30] Y. Li and Y. Yang, “Convergence analysis of two-layer neural networks with relu activation,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 597–607.
[31] B. Liu, M. Wang, H. Forooosh, M. Tappen, and M. Penksy, “Sparse con- volutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 806–814.
[32] W. Wei, C. Wu, Y. Wang, Y. Chen, and L. Hai, “Learning structured sparsity in deep neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2016, pp. 2074–2082.
[33] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (ELUs),” 2015, arXiv:1511.07289. [Online]. Available: https://arxiv.org/abs/1511.07289
[34] H. Ide and T. Kurita, “Improvement of learning for CNN with ReLU activation by sparse regularization,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), May 2017, pp. 2684–2691.
[35] C. Jiang, X. Cheng, H. Gao, X. Zhou, and J. Wan, “Toward compu- tation offloading in edge computing: A survey,” IEEE Access, vol. 7, pp. 131543–131558, 2019.
[36] W.-C. Chen, Y. Xiong, J. Gao, N. Gelfand, and R. Grzeszczuk, “Efficient retraining of robust image features on mobile devices,” in Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality, Nov. 2007, pp. 287–288.
[37] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learn- ing applied to document recognition,” in Proc. Adv. Neural Inf. Process. Syst., vol. 8, pp. 888–895, 1996.
[38] H. Ide and T. Kurita, “Improvement of learning for CNN with ReLU activation by sparse regularization,” in Proc. Int. Joint Conf. Neural Netw. (IJCNN), May 2017, pp. 2684–2691.
[39] C. Jiang, X. Cheng, H. Gao, X. Zhou, and J. Wan, “Toward compu- tation offloading in edge computing: A survey,” IEEE Access, vol. 7, pp. 131543–131558, 2019.
[40] W.-C. Chen, Y. Xiong, J. Gao, N. Gelfand, and R. Grzeszczuk, “Efficient extraction of robust image features on mobile devices,” in Proc. 6th IEEE ACM Int. Symp. Mixed Augmented Reality, Nov. 2007, pp. 287–288.
[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learn- ing applied to document recognition,” in Proc. Adv. Neural Inf. Process. Syst., vol. 8, pp. 888–895, 1996.
[42] Z. He, X. Zhang, Y. Cao, Z. Liu, B. Zhang, and X. Wang, “LiteNet: Lightweight neural network for detecting arrhythmias at resource-constrained mobile devices,” Sensors, vol. 18, no. 4, pp. 1229, Apr. 2018.
[43] Y. Gu, Y. Wang, Z. Liu, J. Liu, and J. Li, “Sleepguardian: An RP-based health-care system guarding your sleep from afar,” IEEE Netw., vol. 34, no. 2, pp. 164–171, 2020.
[44] Y. Cao, F. Wang, X. Lu, N. Lin, B. Zhang, Z. Liu, and S. Sigg, “Contactless body movement recognition during sleep via WiFi signals,” IEEE Internet Things J., vol. 7, no. 3, pp. 2028–2037, Mar. 2020.
[45] Y. Gu, X. Zhang, Z. Liu, and F. Ren, “WiFi-based real-time breathing and heart rate monitoring during sleep,” in Proc. IEEE Global Commun. Conf. (GLOBECOM), Dec. 2019, pp. 1–6.
[46] Y. Gu, C. Zhang, Y. Wang, Z. Liu, Y. Ji, and J. Li, “A contactless and fine- grained sleep monitoring system leveraging WiFi channel response,” in Proc. IEEE Int. Conf. Commun. (ICC), May 2019, pp. 1–5.
[47] J. Yu, J. Li, Z. Yu, and Q. Huang, “Multimodal transformer with multi- view visual representation for image captioning,” IEEE Trans. Circuits Syst. Video Technol., early access, Oct. 15, 2019, doi: 10.1109/TCSVT.2019.2947482.

XINYI LIU received the B.S. degree in computer science from Northwest University, Xi’an, China, in 2019, where she is currently pursuing the mas- ter’s degree. Her main research interests include mobile computing and deep learning.

BAOYING LIU received the Ph.D. degree in cultural heritage conservation technology from Salento University, Italy, in 2012. She is the talent introduced by the Hundred-Talent Program to the School of Information Science and Technology, Northwest University. Her current research interests include wireless networks and the Internet of Things applied in cultural heritage field.
GUOQING LIU received the B.S. degree in mechanical engineering from the Chongqing University of Arts and Science, Chongqing, China, in 2018. He is currently pursuing the master’s degree with Northwest University. His main research interests include deep learning and mobile computing.

FENG CHEN received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xi’an, China, in 2012. His current research interests include localization and performance issues in wireless ad hoc, mesh, and sensor networks.

TIANZHANG XING (Member, IEEE) received the B.E. degree from the School of Telecommunications Engineering and the Ph.D. degree from the School of Information and Technology, Northwest University, Xi’an, China. He is currently an Associate Professor with the School of Information and Technology, Northwest University. His current research interests include mobile computing, pervasive computing and wireless networks, with emphasis on the localization problem and the location-based services.