Edge-Computing System Based on Smart Mat for Sleep Posture Recognition in IoMT

Haikang Diao[1](✉), Chen Chen[2], Xiangyu Liu[3], Amara Amara[4], and Wei Chen[1,2]

[1] Center for Intelligent Medical Electronics, School of Information Science and Technology, Fudan University, Shanghai 200433, China
{19210720023,w_chen}@fudan.edu.cn

[2] Human Phenome Institute, Fudan University, Shanghai 201203, China
chenchen_fd@fudan.edu.cn

[3] College of Communication and Art Design, University of Shanghai for Science and Technology, Shanghai 200093, China

[4] IEEE CASS, Paris, France
amara.amara@tdh.ch

Abstract. Sleep posture has been proven to be a crucial index for sleep monitoring in the Internet of Medical Things (IoMT). In this paper, an edge-computing system based on a smart mat for sleep posture recognition in IoMT is proposed. The system can recognize postures unobtrusively with a dense flexible sensor array. To meet the requirements of embedded system in IoMT, a light-weight algorithm that includes pre-processing, EdgeNet pre-training, model quantization, model deployment is proposed. Finally, the complete algorithm is deployed in embedded systems (STM32) and edge computing for sleep posture monitoring is implement in IoMT. Through a series of short-term and overnight experiments with 21 subjects, results exhibit that the accuracy of the short-term experiment is up to 92.10% and the overnight experiment is up to 75.43%. After quantization, the accuracy of the overnight is up to 74.79%, and the runtime of the complete algorithm is about 65ms in the STM32. Compared with other methods, edge-computing systems have the advantages of low power consumption, low cost, low latency, high reliability, and no risk of privacy leakage. With the promising results, the proposed system is capable of providing sleep posture recognition and can be integrated into IoMT as an edge device.

Keywords: Edge computing · Sleep posture recognition · EdgeNet · Model quantization

1 Introduction

The Internet of Medical Things (IoMT) is a practical application for health care, which combines with the Internet of Things (IoT) devices and MedTech tools [1]. In the field of health care, sleep posture has been proven to be a crucial index for sleep monitoring. Wrong sleep postures may increase the burden of muscles and ligaments and result in

X. Gao et al. (Eds.): MobiHealth 2021, LNICST 440, pp. 85–94, 2022.
https://doi.org/10.1007/978-3-031-06368-8_6

shoulder, neck, or back pain; obstruct the airways to the lung and lead to breathing disorders like sleep apnea; affect the blood circulation, and induce pressure ulcers [2–4]. Therefore, a sleeping posture recognition system that enables long-time sleep posture monitoring and can be integrated into IoMT systems is needed.

Recently, quantities of methods have been proposed to recognize sleep posture. These methods can be classified into three categories. The first is to use video cameras to monitor sleep postures [5], but this method is susceptible to bedsheets and light, and may also produce privacy leaks. The second is to use a variety of wearable sensors to monitor sleep posture [6, 7]. However, with the attached sensors, natural sleep may be disrupted. The third is based on the pressure-sensing smart mat to achieve safe, convenient, comfortable, and non-intrusive sleep posture recognition, which can be applied in hospitals or home scenarios [8, 9]. In our previous study [10], an unobtrusive miniature scale smart mat system based on a dense flexible sensor array along with deep residual networks for sleep posture recognition is proposed. However, all of the methods mentioned above are unable to deploy in embedded systems for edge computing due to the complex algorithms, and therefore cannot be integrated directly into IoMT as an edge device.

In this paper, an edge-computing system based on a smart mat for sleep posture recognition in IoMT is proposed. To meet the requirements of the embedded system, we propose EdgeNet, a highly efficient model, and perform model quantization to significantly compress the model. Finally, the complete algorithm is deployed in embedded systems and edge computing for sleep posture monitoring is implement in IoMT. To verify the feasibility of the proposed system in real scenarios, a series of short-term and overnight experiments and performance evaluations in STM32 were conducted. Compared with other methods, edge-computing systems have the advantages of low power consumption, low cost, low latency, high reliability, and no risk of privacy leakage. The rest of the paper is organized as follows: Sect. 2 describes the system design and implementation. Section 3 presents the algorithm for sleep posture recognition. Section 4 presents the experiment and the results. Section 5 and Section 6 present the discussion and conclusion respectively.

2 System Design and Implementation

The edge computing system is shown in Fig. 1. As introduced in our previous research [10], When a person lies on the mat, the force-sensing mat acquires the voltage in different areas, and then the voltage goes through the signal control and acquisition circuit into the embedded system. Finally, the embedded system is used to convert the voltage signal into a pressure distribution map, and implement the sleep position classification. The electrodes and pins of the mat, the force-sensing resistor, the data acquisition circuit, and the embedded system are shown in Fig. 2.

To implement edge computing, the STM32 system, a cost-effective embedded system, is used as the computing platform. This system is based on STM32F746ZGTx microcontroller with 216 MHz clock speed, 1 Mbyte of Flash memory, and 340 Kbytes of RAM. It provides a lot of peripherals such as serial ports, GPIO, and 12-bit analog-to-digital converter (ADC). Compared to expensive servers, it costs only $5 and can be used as an edge device in IoMT. Meanwhile, it also can run edge algorithms due to its high main frequency and sufficient memory.

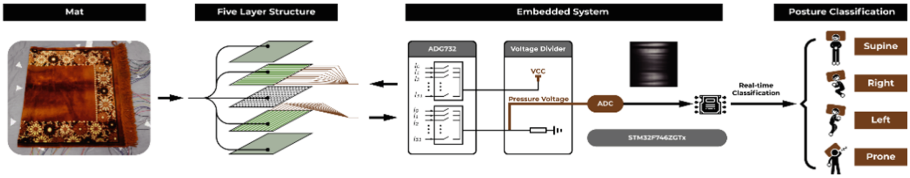
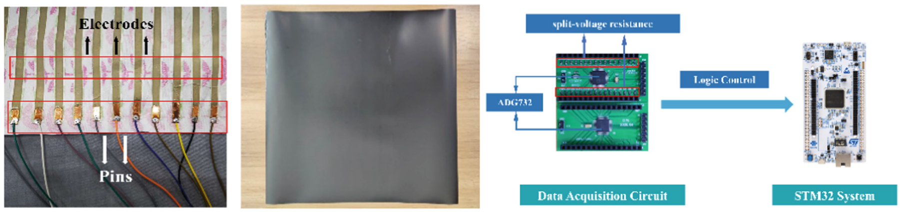


Fig. 1. The structure of the edge computing system.



(a) (b) (c)

Fig. 2. (a) The electrodes and pins of the mat, (b) the force-sensing resistor, (c) the data acquisition circuit, and the embedded system.

The pressure distribution when one hand is pressed on the mat is shown in Fig. 2. Due to the high sensitivity of the system, the position and contour of the palm and fingers can be detected.

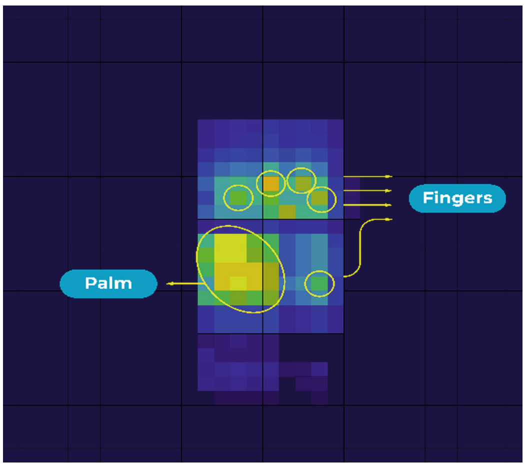


Fig. 3. The pressure distribution when one hand is pressed on the mat.

3 Algorithm Framework

To implement real-time edge computing in IoMT, a light-weight algorithm for sleep posture classification that can be deployed in an embedded system needs to be designed. In this section, an EdgeNet-based sleeping position classification algorithm is proposed as shown in Fig. 3. It consists of four steps: pre-processing, EdgeNet pre-training, model quantization, model deployment. The sleep postures are classified into four categories: supine, prone, right, and left.

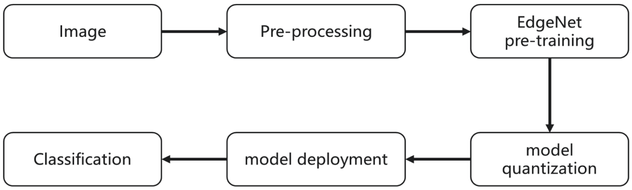


Fig. 4. The overall algorithmic framework.

3.1 Pre-processing

To eliminates the internal noise, high-frequency noise, and redundant information due to the softness and thinness of the pressure sensors embedded in the mat, threshold filtering, Gaussian filtering, and adjacent affected noise removal are used in pre-processing refer to our previous work [10].

3.2 EdgeNet Pre-training

Convolutional Neural Networks (CNN) is a feed-forward neural network, and it has a good effect on image recognition. However, the increase in classification capability comes with another drawback: the size and computational complexity of the model become high and beyond the capabilities of the IoMT system. In this paper, EdgeNet containing 1 Squeeze-and-Excitation block (SE block) [11], 5 bottleneck blocks based on MobileNetV2 [12], and 1 dense layer as shown in Fig. 4. The description of each layer of the network is shown in Table 1: c denotes the number of output channels, s denotes stride, and t denotes expansion factor.

The SE module adaptively reassigns the weights of different feature channels by calculating the interdependencies between channels. MobileNetV2 was proposed by Google Inc in 2018, which is based on the depthwise separable convolutions and inverted residual structure called bottleneck block. The bottleneck block, compare with the normal convolution layer, is more suitable for the embedded system due to the reduction of the FLOPS and parameters.

3.3 Model Quantization

Neural networks are known to be robust to noise and disturbance, and most of the trained weights and activations tend to fall within a small range [13]. Therefore, quantizing the model to reduce the precision requirements for the weights and activations is a convenient and effective way to compress the model. For example, by using 8-bit quantization, we can reduce the size of the model by a factor of 4 with minimum accuracy reduction, which reduces computational complexity, memory consumption, and data access latency.

Table 1. The description of each layer of the network.

Input	Operator	t	c	s
32 × 32 × 1	conv2d	–	16	1
32 × 32 × 16	SE block	–	16	1
32 × 32 × 16	bottleneck	1	8	1
32 × 32 × 8	bottleneck	6	16	2
16 × 16 × 16	bottleneck	6	32	2
8 × 8 × 32	bottleneck	6	64	1
8 × 8 × 64	bottleneck	6	128	2
4 × 4 × 128	avgpool 2 × 2	–	–	–
2 × 2 × 128	FC	–	4	–

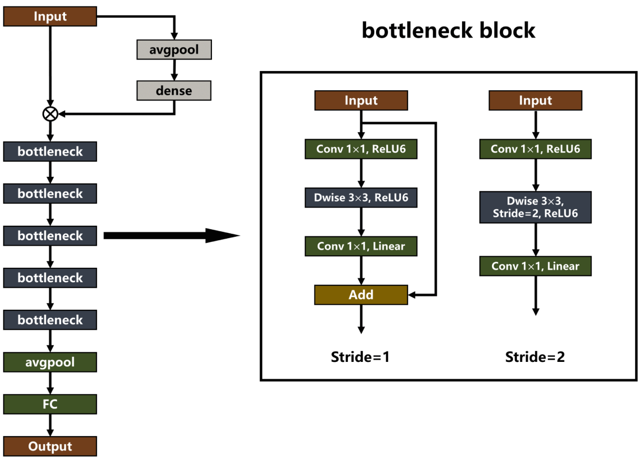


Fig. 5. The structure of the EdgeNet.

In this paper, post-training integer quantization is proposed for model compression after the EdgeNet has been trained. Post-training integer quantization is to convert 32-bit floating-point numbers (such as weights and activation outputs) to the nearest 8-bit fixed-point numbers for the already-trained float model. The operation can be described by the following formula:

$$x_{\text{int}} = \frac{x_f}{s} + z_0 \tag{1}$$

$$s = \frac{f_{\max} - f_{\min}}{2^8 - 1}, z_0 = 255 - \frac{f_{\max}}{s} \tag{2}$$

x_{int} is the quantized value, x_f is the 32-bit value, s denotes the quantized scale and z_0 is the quantized value corresponding to 0.f in a 32-bit value. This is an affine mapping of integers x_{int} to real numbers x_f.

3.4 Model Deployment

The complete sleep posture classification algorithm needs to be deployed to the STM32 to implement edge computing in IoMT. Signal control and pressure data acquisition implemented via the HAL driver library of the STM32. After the neural network model is trained on the TensorFlow platform, it needs to be converted to TF-Lite format before deployment. Then the pre-trained model is converted into an STM32-optimized library by the STM32Cube. AI expansion package. Finally, this STM32-optimized library is integrated into the user project. With this expansion package, neural network models can be inferred in embedded systems instead of on the server.

4 Experiment and Results

In this section, the sleep posture experimental setup, experimental results, and performance evaluation in STM32 are presented.

4.1 Experimental Setup

To evaluate the feasibility of the system for sleep posture recognition, short-term and overnight experiments were conducted as shown in Fig. 5. Sixteen subjects were included in the short-term experiments and five subjects were included in the overnight experiments introduced in our previous research [10]. The mat was placed between the subject's neck and hips and was parallel to the subjects. There were a total of 1059 samples for the short-term experiment and a total of 20521 samples for the overnight experiment.

4.2 Experimental Results

We perform the EdgeNet on the short-term and overnight database respectively. To validate the performance of the EdgeNet, we analyzed the short-term dataset by Leave-One-Person-Out-Cross-Validation. It divides the dataset by subjects, to ensure that the data of the same subject is not included in the training set and validation set at the same time. The overnight database is used as a test set to judge the effectiveness of the proposed algorithm, while the short-term database is used as the training set.

Fig. 6. The video screenshots of the sleep posture collection experiment

For neural network training, the relevant parameters are set as follows: the learning rate is 0.005, the number of iterations for network training is 150, the batch size is 512, the optimization algorithm is adaptive moment estimation (Adam), the loss function is categorical cross-entropy, the momentum is 0.9 and the epsilon is 0.00001 for Batch Normalization.

Table 2 shows the accuracy of the EdgeNet model and other models for sleep posture classification. Compare to other models, the EdgeNet proposed in this paper significantly reduces the size and computational complexity of the model while achieving competitive accuracy. Due to the pre-processing operation, we have simplified the pressure distribution images and removed redundant information, thus achieving high accuracy with the EdgeNet. Meanwhile, by applying the SE module to the input of the network, an attention mechanism is added to the input channels. Thus, the network pays more attention to the channels that are more sensitive to the posture classification task, which improves the efficiency of the model.

Table 2. The model complexity and classification accuracy.

Model	EdgeNet	MobileNet	ResNet10	AlexNet
Short-term	92.10%	93.90%	94.62%	94.90%
Overnight	75.43%	74.82%	77.17%	78.23%
Params	0.11M	2.22M	8.28M	4.37M
Size	0.44 MB	2.49 MB	32.3 MB	17.1 MB
MACC	5.88 M	10.47 M	159.27 M	257.4 M

Table 3. Comparison of EdgeNet before and after quantification

Model	Params	Size	MACC	Overnight
EdgeNet	0.11 M	0.44 MB	5.88 M	75.43%
EdgeNet _Q	0.11 M	0.18 MB	5.94 M	74.79%

It is worth noting that the accuracy of overnight experiments is significantly lower than short-term. First of all, none of the subjects in the overnight sleep experiment had participated in the short-term sleep experiment. Then, to simulate the real sleep situation, subjects may have movements in bed such as turning over as usual. These data are also collected but cannot be effectively identified as to which posture.

Table 3 shows the comparison of the EdgeNet before and after quantization. It can be seen that the size of the model is reduced by 60% after quantization, but the accuracy rate only decreases by 0.64%. This proves that quantization can indeed significantly reduce the model size and computational complexity while ensuring classification accuracy.

4.3 Performance Evaluation in STM32

After the model is quantized, we deploy the EdgeNet in STM32 to evaluate its performance, and for comparison, we also deploy the unquantized EdgeNet in STM32. Table 4 shows the performance and memory usage of the two models in stm32. Since the weights and activation values are quantized from 32 bits to 8 bits, the memory usage, and computational complexity are drastically reduced. Although the MACCs of the two models are similar, the CPU clock cycles required to execute a single operation (Cycle/MACC) are much different. Therefore, the inference time of the quantized model is only 30.7% of that of the unquantized model.

Together with the data acquisition and pre-processing algorithms, the quantized model performs a complete sleep posture acquisition and classification algorithm in less than 65 ms. This proves that the system proposed can be used as an edge device to achieve long-time sleep posture monitoring in IoMT.

Table 4. Performance evaluation in STM32

Model	ROM	RAM	Activations	Runtime	MACC	Cycle/ MACC	Accuracy
EdgeNet	425 KB	228 KB	224 KB	204.58 ms	5.88 M	7.552	75.43%
EdgeNet_Q	111 KB	132KB	128 KB	62.84 ms	5.94 M	2.28	74.79%

5 Discussion

The key index to implement edge computing in IoMT is how to deploy algorithms on small memory, low performance, low power embedded systems, and ensure the utility of the algorithms (low latency and high accuracy). Therefore, the edge algorithm needs to ensure low memory usage and low computational complexity. In our method, compared to the ordinary convolutional layer, bottleneck block can express richer features by expansion factor while reducing memory usage and the number of calculations. Meanwhile, the memory accesses are 28% less than that of the ordinary convolutional layer, which is related to the model runtime.

The mat proposed can still be enhanced by involving the following points. Currently, the model must be trained on the server first and then deployed to the embedded system.

If the model training can be implemented in the embedded system, then the model can be updated and improved according to the usage scenarios. What's more, the respiration rate detection will be included to extend the functionality of the system.

6 Conclusion

In this paper, for offering an edge-computing system for sleep posture recognition in IoMT, a smart mat system based on the EdgeNet is proposed. The algorithmic framework consists of pre-processing, EdgeNet pre-training, model quantization, and model deployment. Experimental results show that the accuracy of the short-term experiment is up to 92.10%, and the accuracy of the overnight is up to 75.43% before quantization. After quantization, the accuracy of the overnight is up to 74.79%, and the runtime of the complete algorithm is about 65ms in the STM32 system, which shows that its ability to provide sleep posture recognition in IoMT. Compare to other methods, our system has the advantages of low cost, low latency, low power consumption, high reliability and no privacy concerns. In the future, the mat system can still be enhanced by integrating embedded network training algorithms and breathing detection.

Acknowledgment. This work was supported in part by the Shanghai Committee of Science and Technology under Grant No.20S31903900, in part by Shanghai Municipal Science and Technology Major Project under Grant No. 2017SHZDZX01, in part by Shanghai Municipal Science and Technology International R&D Collaboration Project (Grant No. 20510710500), and in part by the National Natural Science Foundation of China under Grant No.62001118.

References

1. Ghubaish, A., Salman, T., Zolanvari, M., Unal, D., Al-Ali, A.K., Jain, R.: Recent advances in the internet of medical things (IoMT) systems security. In IEEE Internet Things J. **8**(11), 8707–8718 (2021). https://doi.org/10.1109/JIOT.2020.3045653
2. Parish, J.M.: Sleep-related problems in common medical conditions. Chest **135**(2), 563–572 (2009)
3. Winsky-Sommerer, R., de Oliveira, P., Loomis, S., Wafford, K., Dijk, D.J., Gilmour, G.: Disturbances of sleep quality, timing and structure and their relationship with other neuropsychiatric symptoms in Alzheimer's disease and schizophrenia: insights from studies in patient populations and animal models. Neurosci. Biobehav. Rev. **97**, 112–137 (2019)
4. Soban, L.M., Hempel, S., Munjas, B.A., Miles, J., Rubenstein, L.V.: Preventing pressure ulcers in hospitals: a systematic review of nurse-focused quality improvement interventions. Jt. Comm. J. Qual. Patient Saf. **37**(6), 245-AP16 (2011)
5. Nuksawn, L., Nantajeewarawat, E., Thiemjarus, S.: Real-time sensor- and camera-based logging of sleep postures. In: 2015 International Computer Science and Engineering Conference (ICSEC), pp. 1–6 (2015)
6. Chang, M., et al.: Multimodal sensor system for pressure ulcer wound assessment and care. IEEE Trans. Ind. Inform. **14**(3), 1186–1196 (2018)
7. Jiang, P., Zhu, R.: Dual tri-axis accelerometers for monitoring physiological parameters of human body in sleep. In: 2016 IEEE Sensors, Orlando, pp. 1–3 (2016)

8. Matar, G., Lina, J.-M., Kaddoum, G.: Artificial neural network for in-bed posture classification using bed-sheet pressure sensors. IEEE J. Biomed. Health Inform. **24**(1), 101–110 (2020)
9. Xu, X., Lin, F., Wang, A., Hu, Y., Huang, M.C., Xu, W.: Body-Earth mover's distance: a matching-based approach for sleep posture recognition. IEEE Trans. Biomed. Circuits Syst. **10**(5), 1023–1035 (2016)
10. Diao, H., et al.: Deep residual networks for sleep posture recognition with unobtrusive miniature scale smart mat system. IEEE Trans. Biomed. Circuits Syst. **15**(1), 111–121 (2021). https://doi.org/10.1109/TBCAS.2021.3053602
11. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 8, pp. 2011–2023, 01 August 2020
12. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.:MobileNetV2: inverted residuals and linear bottlenecks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4510-4520 (2018). https://doi.org/10.1109/CVPR.2018.00474.
13. Jacob, B., et al.: Quantization and training of neural networks for efficient integer-arithmetic-only inference. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, pp. 2704–2713 (2018)