AN AUTOMATED DETECTION SYSTEM FOR ASPIRATION DURING SLEEP BASED ON DEEP LEARNING

KENTARO MORI[1], YUTAKA HATA[2], TOSHIYUKI SAWAYAMA[3], YASUMITSU FUJII[4], YOSHIAKI SAJI[4], YOSHITADA SAKAI[5], NAOMI YAGI[6]

[1] Electrical and Computer Engineering, National Institute of Technology (KOSEN), Maizuru College, Kyoto, Japan
[2] Graduate School of Information Science, University of Hyogo, Hyogo, Japan
[3] NEW SENSOR Incorporated, Hyogo, Japan
[4] Ishikawa Hospital, Hyogo, Japan
[5] Division of Rehabilitation Medicine, Kobe University Graduate School of Medicine, Hyogo, Japan
[6] Advanced Medical Engineering Research Institute, University of Hyogo, Hyogo, Japan
E-MAIL: k.mori@maizuru-ct.ac.jp

Abstract:

Aspiration pneumonia has high incidence and mortality rates among the elderly. Continuous patient monitoring is essential for prevention, but it remains challenging. This study proposes a system to predict aspiration during sleep based on deep learning. The system captures facial images of sleeping subjects using infrared cameras. Time series feature values based on the facial action coding system are computed using the artificial intelligence library MediaPipe. The deep learning model composed by a convolutional neural network and a long short-term memory model detects aspiration during sleep from the feature values. We conducted an experiment on nine healthy subjects to obtain facial image data. We used these images to train the model and evaluate its prediction accuracy. The results presented that the model could predict two types, normal and pained expressions, with high accuracy. This work is expected to improve the safety of elderly people and reduce the workload of health and care professionals.

Keywords:

Aspiration; Convolutional neural network; Facial action coding system; Facial expression recognition; long short-term memory; MediaPipe; Pneumonia;

1. Introduction

Pneumonia is one of the common causes of death. The World Health Organization (WHO) reported that lower respiratory infections, including pneumonia, were the fifth most common cause of death worldwide in 2021 [1]. Based on data from the Global Burden of Disease [2], Our World in Data reported that the highest death rates from pneumonia occurred in people aged 70 and older [3]. The incidence of aspiration pneumonia is extremely high in the elderly.

Swallowing is a series of actions that move food from the pharynx to the esophagus and stomach. Dysphagia is a condition characterized by difficulty swallowing. Aspiration refers to the misdirection of food or liquid into the airway. The number of people aged 65 and older in Japan has reached a record high of 34.45 million [4], accounting for 27.6% of the total population, approximately one in four people. In addition, an interview-based survey found that approximately one-third of people aged 65 and older have dysphagia [5]. Therefore, it is estimated that approximately 10 million seniors in Japan have dysphagia. More than 90% of pneumonia-related deaths occur in people aged 65 and older. With the increasing number of pneumonia-related deaths due to the aging population, the development of an aspiration detection system is an urgent issue. It has also been reported that choking accidents due to aspiration are more common in the elderly [6]. When aspiration causes a serious condition accompanied by severe pain, such as choking, signs such as (1) a pained expression and (2) the universal sign for choking (placing the hands on the throat) will appear. Facial expressions are considered to reflect a patient’s psychological state and various physical conditions.

In recent years, deep learning has been increasingly applied to the medical field. For example, some studies have reported feature analysis of medical images using convolutional neural networks (CNNs) [7], and time series analysis using a hybrid model combining CNNs and long short-term memory (LSTM) [8]. In addition, evaluation of swallowing function using an information science approach has been reported [9].

In this study, we develop a system to automatically detect aspiration during sleep and its precursors by using deep learning. In healthcare settings, many patients with pneumonia are hospitalized. To prevent harm caused by aspiration, nurses and caregivers need to monitor the patients during hospitalization. However, there are problems such as the inability to perform constant monitoring, limitations in instant response, and the significant amount of labor required. Therefore, we aim to reduce the harm and the burden associated with aspiration by developing the automatic detection system. The system captures the patient's facial expressions by cameras. Pained expressions caused by aspiration is detected using facial expression recognition technology based on deep learning. This approach enables the automatic detection of aspiration.

2. Automated Detection System

The proposed system monitors subjects using a monitoring system. Feature values are extracted from images captured by the monitoring system. A deep learning model detects pained expressions based on the feature values. In this section, each module of the proposed system is described in detail.

2.1. Monitoring System

We use two infrared cameras to monitor the subjects, because these cameras can also operate at night when people are sleeping. The infrared cameras are positioned in two directions to capture different sleeping positions, such as supine, left lateral, and right lateral. They recorded images of the facial area. Figures 1 and 2 show the placement of the infrared cameras and an actual photograph of the system.

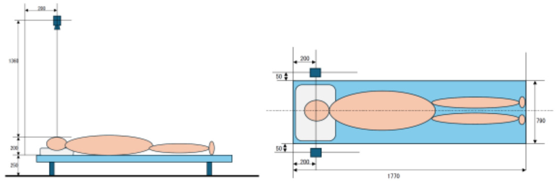


FIGURE 1. Placement of infrared cameras

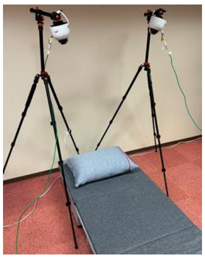


FIGURE 2. Actual photograph of system

2.2. Extraction of Feature Values

To recognize facial expressions, we employ MediaPipe's face landmark detection [10, 11]. This framework estimates a 3D face mesh from a 2D image by placing facial landmarks. The mesh, which represents a curved surface composed of numerous polygons, is applied to the face. The shape of the face is estimated by computing the coordinates of the mesh vertices. A total of 478 3D face landmarks are estimated by using MediaPipe.

Siam et al. [12] defined ten types of feature values for detecting emotions from human facial images, based on landmarks obtained by MediaPipe. In our study, these values are employed to detect pained expressions caused by aspiration. The feature values are computed following the procedure described in reference [12]. A total of 27 key landmarks, based on the Facial Action Coding System (FACS) [13], are selected from the 478 landmarks provided by MediaPipe, as shown in Table 1. An emotion mesh is created by connecting these landmarks, as shown in Table 2. Ten pairs of angles between edges in the emotion mesh, as listed in Table 3, are computed and used as feature values. The system employs these ten values as features to detect pained expressions caused by aspiration.

TABLE 1. Key landmarks [12]

Key landmark ID	MediaPipe landmark ID	Description
0	61	Mouth end (right)
1	292	Mouth end (left)
2	0	Upper lip (middle)
3	17	Lower lip (middle)
4	50	Right cheek
5	280	Left cheek
6	48	Nose right end
7	4	Nose tip
8	289	Nose left end
9	206	Upper jaw (right)
10	426	Upper jaw (left)
11	133	Right eye (inner)
12	130	Right eye (outer)
13	159	Right upper eyelid (middle)
14	145	Right lower eyelid (middle)
15	362	Left eye (inner)
16	359	Left eye (outer)
17	386	Left upper eyelid (middle)
18	374	Left lower eyelid (middle)
19	122	Nose bridge (right)
20	351	Nose bridge (left)
21	46	Right eyebrow (outer)
22	105	Right eyebrow (middle)
23	107	Right eyebrow (inner)
24	276	Left eyebrow (outer)
25	334	Left eyebrow (middle)
26	336	Left eyebrow (inner)

TABLE 2. Emotion mesh [12]

Edge	Connected vertices IDs	Edge	Connected vertices IDs
1	(0, 2)	20	(23, 22)
2	(0, 3)	21	(22, 21)
3	(1, 2)	22	(21, 12)
4	(1, 3)	23	(12, 13)
5	(7, 6)	24	(12, 14)
6	(7, 8)	25	(11, 13)
7	(6, 4)	26	(7, 19)
8	(8, 5)	27	(11, 14)
9	(6, 9)	28	(14, 4)
10	(9, 0)	29	(20, 26)
11	(4, 0)	30	(26, 25)
12	(8, 10)	31	(25, 24)
13	(10, 1)	32	(24, 16)
14	(7, 19)	33	(16, 17)
15	(7, 20)	34	(16, 18)
16	(7, 0)	35	(15, 17)
17	(7, 1)	36	(15, 18)
18	(19, 23)	37	(18, 20)
19	(19, 14)	38	(18, 5)

TABLE 3. Feature values [12]

Feature	Enclosing vertices IDs
θ_1	(2, 0, 3)
θ_2	(0, 2, 1)
θ_3	(6, 7, 8)
θ_4	(9, 7, 10)
θ_5	(0, 7, 1)
θ_6	(1, 5, 8)
θ_7	(1, 10, 8)
θ_8	(13, 12, 14)
θ_9	(21, 22, 23)
θ_{10}	(6, 19, 23)

2.3. Detection Model for Pained Expressions

A deep learning model is employed to detect pained expressions from the feature values. Figure 3 shows the architecture of the model. The model, which combines a CNN and an LSTM, is expected to achieve high prediction accuracy on time-series data. The model input consists of 40 time steps of the feature values, while the output consists of a binary classification: "pained expression" or "normal expression."

3. Dataset

We conducted an experiment on nine healthy subjects to obtain pained expression data by using the monitoring system. We employed an 8MP dome-type IP camera (DH-IPC-HDBW2841RN-ZAS, WDJ Hi-Tech Inc.) as the infrared camera. The sampling rate was 20 fps, and the image size was 2,160 pixels vertically and 3,840 pixels horizontally. The subjects' facial expressions were recorded for approximately 30 seconds using two infrared cameras installed on the left and right sides of the subjects, according to the procedure described in Table 4. The experiments were conducted under both lighting and non-lighting conditions. A total of 18 samples were obtained from nine subjects. In this study, time periods annotated as "pained expression" were labeled as accordingly, and all remaining periods were labeled as "normal expression." After applying MediaPipe to the images, the feature values were computed and subsequently normalized by dividing them by 180 degrees. The feature values were then smoothed using a moving average. After smoothing, the average of each feature value was shifted to zero. Feature values from the left and right cameras were combined by computing their average. If a feature value was missing from one side, it was replaced with the corresponding value from the other side.

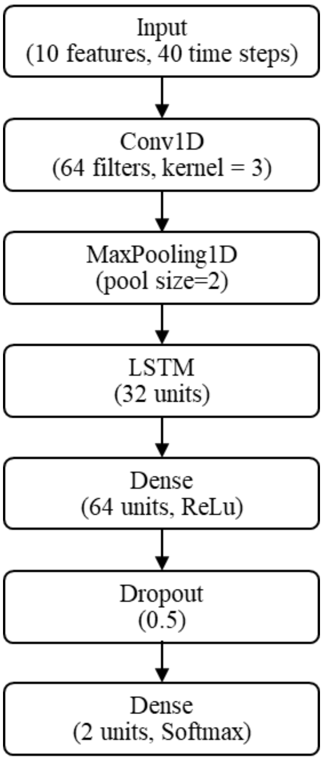


FIGURE 3. Architecture of model

TABLE 4. Procedure of experiment

Step	Facial expression
1	Normal expression [15 sec]
2	Pained expression [15 sec]

4. Evaluation Experiments

The dataset was used for training and testing the detection model. To evaluate the model, data from a single subject were used as the test set, and data from the remaining subjects were used for training. We evaluated the model by iteratively using each subject’s data as the test set and training on the data for the remaining subjects. The dataset was constructed by extracting 40 frames of data on the time series feature values, shifting each frame by one frame. The number of data samples was increased by adding noise and scaling the training data. This model was trained using the Adam optimizer (learning rate: 0.00001) and the binary cross entropy was used as the loss function. Training was performed for 20 epochs with a batch size of 256.

5. Results

Table 5 shows the confusion matrix. Accuracy, recall, precision and F1-score are shown below the table. The values are calculated as follows.

Accuracy: (TP+TN) / (TP+FP+FN+TN)
Recall: TP / (TP+FN)
Precision: TP / (TP+FP)
F1-score: (2×Precision×Recall) / (Precision+Recall)

The model achieved an accuracy of 0.87, a recall of 0.94, a precision of 0.85, and an F1-score of 0.89. Figures 4(a), 4(b), and 4(c) show visualizations of the prediction results. Examples of high-, medium-, and low-quality prediction results are shown in Figures 4(a), 4(b), and 4(c), respectively. The horizontal axis of each graph shows the elapsed time, and the color of the graph shows the state. Red regions show the “pained expression,” and blue regions show the “normal expression.” The upper part of each graph shows the true labels, and the lower part shows the prediction results by the models. As shown in Table 5 and Figure 4, the model performs well overall, but a few misclassifications are observed.

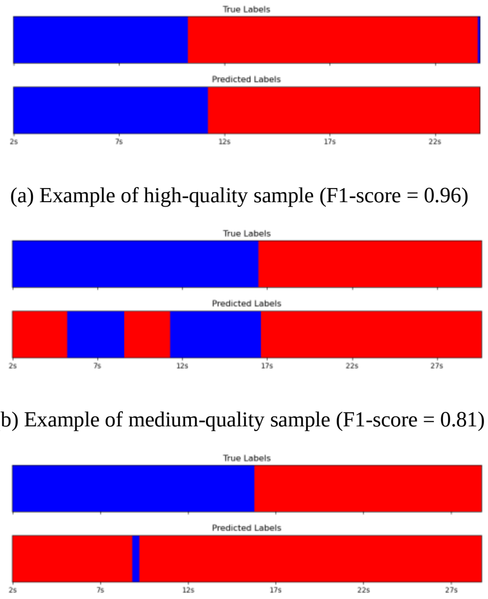
6. Discussion

According to the evaluation experiments, we confirmed that the model correctly predicted facial states for most of the data. However, false detections occurred in some of the data, as shown in Figure 4 (c). Figures 5 and 6 show the learning curves for the model loss. Figure 5 shows the average curves across all models, while Figure 6 shows the curves for the individual model that produced the low-quality sample. The blue line shows the loss for the training data, and the orange line shows the loss for the test data. As shown in Figure 5, the loss on the test data converges. In contrast, Figure 6 shows divergence. This result shows that overfitting has occurred on some of the data. In this study, we used only a small amount of simple data. To solve the problem of overfitting, it is necessary to create large-scale datasets with variations.

TABLE 5. Confusion matrix

		Predicted class	
		Pained	Normal
Actual class	Pained	TP: 4,799	FN: 328
	Normal	FP: 867	TN: 3,559

Accuracy: 0.87, Recall: 0.94, Precision: 0.85, F1: 0.89



(a) Example of high-quality sample (F1-score = 0.96)

(b) Example of medium-quality sample (F1-score = 0.81)

(c) Example of low-quality sample (F1-score = 0.66)

FIGURE 4. Visualizations of prediction results

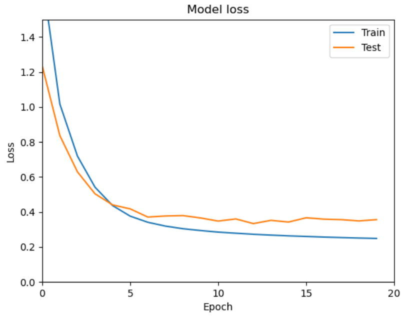


FIGURE 5. Average learning curves of all models

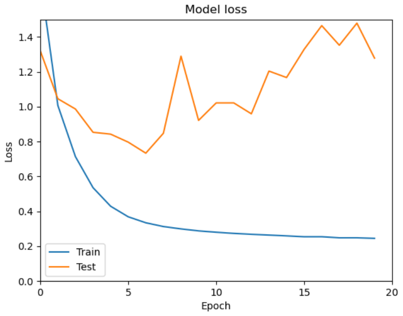


FIGURE 6. Learning curves of individual model (low-quality sample)

7. Conclusion

In this study, we proposed a system to predict aspiration during sleep based on deep learning. Facial images of the sleeping subject were obtained using infrared cameras. Time series feature values based on the FACS were computed using the artificial intelligence library MediaPipe. The deep learning model, which combines a CNN and an LSTM, predicted aspiration during sleep from the feature values. In the evaluation experiments, we confirmed that the model could predict two types, normal and pained expressions, with high accuracy. However, some models suffered from overfitting due to the small size of the dataset. To address this problem, the model needs to be trained with more diverse datasets. Since the proposed method is based on deep learning technology, it is possible to develop a robust model by collecting diverse datasets. Therefore, if such data can be collected, the system could be effectively implemented. We believe that the system will contribute to ensuring the safety and well-being of the elderly individuals and help reduce the burden on healthcare and nursing staff.

Acknowledgements

This work was supported by Himeji City Industry-Academia Collaborative Research Grant.

References

[1] World Health Organization (WHO), The top 10 causes of death, https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] Global Burden of Disease Collaborative Network, Global Burden of Disease Study 2021 (GBD 2021), Seattle, United States, Institute for Health Metrics and Evaluation (IHME), 2024.

[3] Our World in Data, Pneumonia, https://ourworldindata.org/pneumonia?utm_source=chatgpt.com

[4] Statistical Handbook of Japan https://www.stat.go.jp/index.html

[5] N. Roy, J. Stemple, R. M. Merrill, and L. Thomas, "Dysphagia in the elderly: preliminary evidence of prevalence, risk factors, and socioemotional effects Affiliations expand", Annals of Otology, Rhinology & Laryngology, Vol. 116, No. 11, pp.858-865, Nov. 2007.

[6] Y. Taniguchi, M. Iwagami, N. Sakata, T. Watanabe, K. Abe, and, N. Tamiya, "Epidemiology of Food Choking Deaths in Japan: Time Trends and Regional Variations", Journal of Epidemiology, Vol. 31, No. 5, pp.356-360, May 2021.

[7] K. Mori, K. Kitaya, T. Ishikawa, and Y. Hata, "A Pregnancy Prediction System based on Uterine Peristalsis from Ultrasonic Images", Intelligent Automation & Soft Computing, Vol. 29, No. 2, pp. 335-352, Jun. 2021.

[8] K. Mori, Y. Tange, Y. Adachi, and T. Gonda, "Deep Temperature Estimation for Hyperthermia Therapy Based on Surface Temperature and Ultrasonic Image Using Deep Learning", ICIC Express Letters, Part B: Applications, Vol. 15, No. 8, pp.865-872, Aug. 2024.

[9] A. Yoshida, N. Yagi, Y. Fujii, H. Shibutani, Y. Kobayashi, Y. Saji, Y. Sakai, and Y. Hata, "A Visible Camera Approach to Motion Tracking based Swallowing Evaluation", Proceeding of ICMLC 2024, Sep. 2024.

[10] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.L. Chang, M. Yong, J. Lee, W.T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Perceiving and Processing Reality", Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019.

[11] Google AI for Developers https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker

[12] A. I. Siam, N. F. Soliman, A. D. Algarni, F. E. Abd El-Samie, and A. Sedik, "Deploying Machine Learning Techniques for Human Emotion Detection", Computer Intelligence and Neuroscience, Vol. 2022, No. 1, 2022.

[13] P. Ekman, "Universals and cultural differences in facial expressions of emotion", Nebraska Symposium on Motivation, Vol. 19, pp.207–283, 1971.