데이터사이언스학 석사학위 논문

Non-contact Vision-based Obstructive Sleep Apnea Detection using Infrared Sleep Video

적외선 수면 동영상을 활용한 비접촉 비전 기반 폐쇄성 수면 무호흡증 진단

2023년 2월

서울대학교 대학원

데이터사이언스학과

어 경 선

Non-contact Vision-based Obstructive Sleep Apnea Detection using Infrared Sleep Video

적외선 수면 동영상을 활용한 비접촉 비전 기반 폐쇄성 수면 무호흡증 진단

지도교수 김 형 신

이 논문을 데이터사이언스학 석사 학위논문으로 제출함

2022년 12월

서울대학교 대학원

데이터사이언스학과

어 경 선

어경선의 데이터사이언스학 석사 학위논문을 인준함

2023년 1월

위 원 장 Wen-Syan Li (인)

부위원장 김 형 신 (인)

위 원 오 민 환 (인)

Abstract

Gyeongseon Eo
Data Science Major
Department of Data Science
Graduate School of Data Science
Seoul National University

We present a contactless vision-based obstructive sleep apnea (OSA) detecting method that can achieve results quickly and comfortably. To this end, three approaches are taken. First, a new dataset is constructed around events, away from epoch, which is the basic analysis unit of sleep research. Second, an attempt is made to utilize respiratory arousal to detect OSA. Finally, in order to reduce the amount of computation of the model, the difference in pixelwise values between frames is used without using the optical flow. In addition, a robust model is created using 617 sleep data, which is several times more than previous studies. As a result, we achieved 74% accuracy with f1 score of 0.84 while having 95% fewer flops compared to the baseline model.

Keywords: Obstructive Sleep Apnea (OSA), Respiratory Arousal, Apnea-Hypopnea Index (AHI), Video Classification

Student ID: 2021-24878

Contents

List of Figures

List of Tables

1 Introduction

For modern people who are getting busier, recovering from fatigue through sleep is a key element in health. Despite its importance, however, the number of sleep disorders in modern people is increasing. Among them, the representative sleep disease is obstructive sleep apnea (OSA). OSA refers to a repeated symptoms in which breathing is temporarily stopped due to blocked air flow in the respiratory tract during sleep. If a person suffer from OSA, snoring, apnea, and respiratory arousal continue to appear, making it difficult to fall into deep sleep and lowering the quality of sleep. According to a recent study, obstructive sleep apnea is suspected in about 1 in 6 adults in Korea [23], so it can now be seen as one of the major health concerns. However, at present, the only accurate way to determine whether or not a person suffers from sleep apnea is to take a polysomnography (PSG). Polysomnography refers to a examination that is received while sleeping overnight in a hospital with various sensor devices (brain waves, eye movements, breathing movement, etc.) attached to the body. When sleep is over, an expert reads the results from the sensors, and the doctor checks the results again to determine whether there is a sleep disorder. The examination measures the average number of apneas or hypopneas that occur per hour during sleep (apnea-hypopnea index (AHI)). If an AHI is 5 or more, sleep apnea is diagnosed. The AHI of 15 or more is considered moderate, and the AHI of 30 or more is considered severe sleep apnea. This PSG has the disadvantage of being complicated to perform and that it may show a different sleep pattern than usual because subject sleep in a sleep environment different from home, and it requires a lot of manpower and facilities, so it is economically expensive and takes a long time to interpret. Therefore, it is not easy for everyone to go for a PSG with a light heart. Here, if OSA can be detected without PSG and the detection method is easily accessible to anyone in everyday life, it will be of great help to many people suffering from sleep disorders.

An easy and fast way to detect obstructive sleep apnea without sensing results of PSG is to use an AI model to analyze sleep videos to determine whether or not subject have OSA. In fact, when taking PSG, sleep videos during the examination are recorded using an infrared camera, so the videos can be used as data to create a deep learning model. If OSA can be detected only with sleep videos, there is no need to sleep inconveniently with the sensor attached, and there is no need to visit a hospital and go to bed in an unfamiliar environment. In addition, if sleep video analysis can be performed on edge devices such as portable medical devices or smartphones, accessibility to examinations will be further improved. We will introduce a good performance but lightweight OSA detecting system as described above.

2 Related Work

There have been steady attempts to detect OSA using deep learning models. The majority of studies utilize bio-signals, which have taken appropriate preprocessing and classified the results through deep learning models. The most frequently used signal was electrocardiogram (ECG) measured through PSG [28, 21, 27, 4, 24, 10, 1], followed by electroencephalogram (EEG) [1, 5, 25, 17], There have also been studies using electromyography (EMG)[3, 18]. In addition to bio-signals obtained through PSG, there have been studies using pulse oximetry signals (SpO2) for detection [19, 14, 11]. Unlike PSG, these studies are characterized by using only a single channel for detection, and mainly used CNN or LSTM architecture. The detecting performance is excellent, with an accuracy of 88% to 98%. However, this signal-based detection has the limitation of still having to go to bed with the sensor attached to the body to measure the signal even if it is a single channel. And the need to have a device to measure the signal can also be seen as another disadvantage.

As well as bio-signal data, there are also non-contact detecting studies using audio signals[7, 12]. In the case of using sound, OSA was classified using respiratory sound signals or snoring sounds, and performance was about accuracy of 90% to 95%. Sound is non-contact and has the advantage of requiring only a recorder, but has the disadvantage of being easily mixed with noise. In a controlled environment such as a hospital, it is possible to record the sound of breathing or snoring only, but in a place such as a house, there is a high possibility that living noises are mixed in, resulting in poor generalization performance.

Lastly, some studies have attempted to detect OSA using sleep videos. In the case of [26], a rule-based method was used instead of machine learning, and motion was detected by utilizing the difference in intensity between frames. Although it showed 94% detection accuracy, there is a limit to the rule-based structure in which eight parameters must be found each time for detection. [29] used a random forest structure to predict AHI and classify whether or not the AHI was 15 or less, but the accuracy was not as good as 74%. [2] showed classification accuracy of 83% by OSA classification and estimating AHI models using dense optical flow and 3D CNN structure. However, if we verify at the video data used in the experiment, first of all, there is a problem that the reality is low because the subject are sleeping without covering the blanket and the video was recorded with the camera positioned right above the body. In addition, because the model structure is too heavy, it takes up to 20 hours to inference a 5 hour sleep video.

Furthermore, a common limitation of all of the above studies is that the number of cases of data used is too small to generalize. Most of the studies were conducted with only 30 to 70 sleep cases, with at least 15 people and at most 150 people.

3 Materials and Method

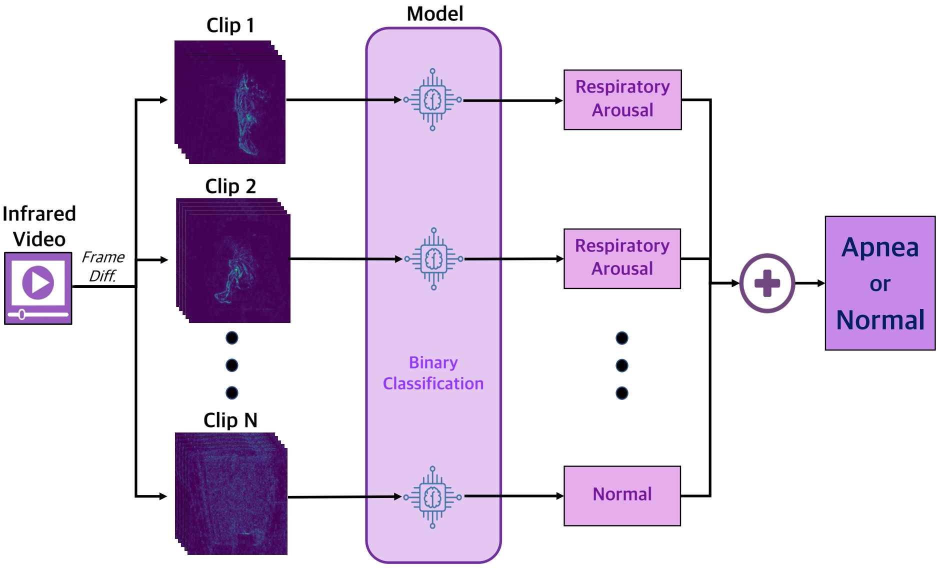

Figure 3.1 OSA detecting system using infrared sleep videos

In this section, a system for detecting obstructive sleep apnea through sleep videos will be introduced. The whole process is depicted in Figure 3.1.

3.1 Data Description

In this study, Infrared Sleep Video Data for Diagnosing Sleep Disorder dataset provided by AI Hub [13] was used. This dataset consists of 1000 infrared sleep videos taken together with PSG performed for 4 years in 4 hospitals (A to D). Each video is an mp4 file of about 6 hours, with a resolution of 640x480 and a frame rate of about 5 fps. Using the PSG results, annotation files labeled according to the time of each event in the video is provided. This event includes various information such as sleep stage, apnea, hypopnea, respiratory arousal, and snoring. Event labeling basically records sleep stages (Wake, N1, N2, N3 and REM) in unit of an 30-second epoch, and when sleep disturbance events such as apnea and respiratory arousal occur, the start and end times of the event are written. Looking at the sleeping state in the video, as shown in Figure 3.2, the face of a sleeping person is mosaic-processed and video was taken from a ceiling angle while sleeping with a blanket on. And if the hospital that collects the data is different, the equipment that films the videos, the machine that performs the PSG are different, and the person who reads the results is also different.

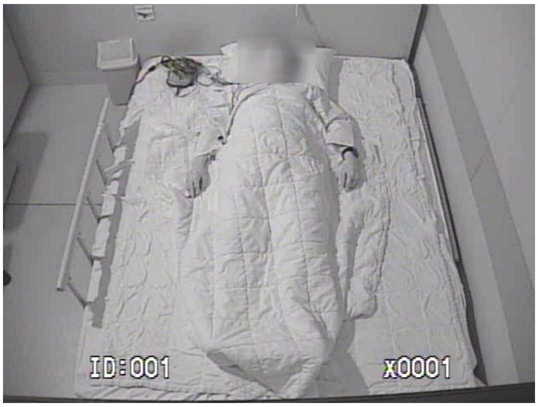

Figure 3.2 Sleep video frame example

In order to detect OSA through videos, it is necessary to observe the sleep pattern of the subject of PSG to determine whether the subject is currently sleeping normally or having apnea-hypopnea. If we look at the video data of subjects with OSA, they sleep breathing normally, but when they enter obstructive sleep apnea, breathing literally stops temporarily and there is little movement of the abdomen. Then, oxygen saturation decreases and respiratory arousal is induced by feeling difficulty in breathing. When respiratory arousal outbreak, the subject breathe rapidly, and at this time, large movements appear throughout the body, including the abdomen. The shorter the cycle repeats, the more severe OSA patients become. And those symptoms can occasionally appear in people who are not diagnosed with OSA. As such, frequency differences are an important factor in detecting OSA. In order for the deep learning model to train and classify the corresponding aspects in the video, it is possible to set training data by labeling apnea-hypopnea as a positive class and normal breathing as a negative class. However, as mentioned above, the difference in movement change between normal breathing and respiratory arousal states is much larger than that between normal breathing and apnea-hypopnea states. Therefore, for the purpose of this study to use the vision-based algorithm, it would be easier to distinguish the state of respiratory arousal as positive and normal breathing as negative. However, in order to do this, it is necessary to establish grounds that OSA can be detected only by distinguishing whether there is respiratory arousal rather than whether or not apnea-hypopnea outbreak. According to the current standard, OSA diagnosis is performed by measuring the AHI level through the PSG examination, and based on the level, OSA status and severity are determined. In other words, if AHI can be predicted only by presence or absence of respiratory arousal, this is in line with OSA detection. Figure 3.3 is a chart showing the relationship between the number of occurrences of respiratory arousal per sleep of subjects and AHI. As can be seen in the chart, since the number of respiratory arousal and AHI have an almost linear relationship, if the number of respiratory arousal can be accurately predicted, the AHI also can be predicted using this relationship, and eventually OSA can be detected.

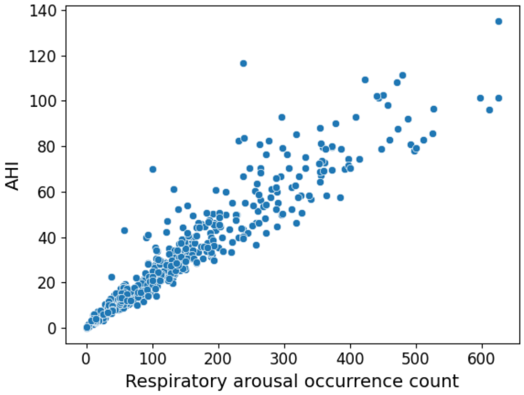

Figure 3.3 Number of occurrences of respiratory arousal per sleep and AHI

However, there is an issue with the dataset. Since apnea-hypopnea accompanies respiratory arousal, there should not be a significant difference in the ratio between the two events, but some hospitals that collected data had problems with labels. In the case of Hospital A and Hospital D, the ratio of the number of respiratory arousal event labels to the number of apnea-hypopnea event labels was similar at about 1:1.1, but in the case of Hospital B and Hospital C, the ratio was 1:6.76 and 1:0.73, respectively. Therefore, only data from 499 of Hospital A and 118 of Hospital D, which were judged to have no problem in terms of ratio, were used in our study. In the case of Hospital A and Hospital D, the ratio between events is not exactly 1:1, and the reason why there are slightly more apnea-hypopnea events is that the respiratory arousal duration is proportional to the immediately preceding apnea-hypopnea duration, but hypopnea or short apnea may not cause respiratory arousal. Of the 499 cases of Hospital A data, 399 cases (about 80%) were divided into training data, 10% (50 cases) as validation data, and the remaining 10% as test data. The 499 cases were not randomly split according to the ratio, but were divided so that people with low to high AHI values were evenly included in each dataset. The 499 cases were lined up according to the AHI values and split by a method of selecting with a constant step. In the case of Hospital D, all cases were used as test data. The reason why all of the Hospital D data were included in the test data instead of using some of them in the training data is to verify the generalization performance and whether it is possible to classify data collected from medical institutions in other environments by training only the data of Hospital A.

3.2 Detection of Respiratory Arousal within Video Clips

Creating Video Clips The biggest problem that arises when constructing training data to detect OSA using the Infrared Sleep Video Data for Diagnosing Sleep Disorder dataset is to divide the data into a form suitable for the video model. Each sleep video consists of about 6 hours, and it is impossible to put it into the model at once, so we have to divide it into clips of an appropriate length according to the size of the model. But the problem is that there are no general rules here. First of all, looking at the part that determines the clip length, in the case of general sleep research, work is done in sequential epoch units from the start to the end of sleep. Here, an epoch is a standard division unit for sleep analysis, and one epoch is 30 seconds. For example, in the case of sleep stage classification, it is divided into epoch units. However, the current goal is not to classify sleep stages in clips, but to determine whether respiratory arousal outbreak in clips, so the clip length is not necessarily set at 30 seconds. And not only the length of the clip, but also the labeling of each clip causes problems. For example, if we decide to use clips in sequential 30-second epoch units, we should label each epoch whether it is normal breathing or respiratory arousal. At this moment, the criterion for labeling as positive becomes ambiguous. This is because it is unclear whether to judge positive when an event lasts for more than a few seconds out of 30 seconds. In the case of respiratory arousal, since it is an event that occurs for 3 seconds if it is short and 30 seconds or more if it is long, the labeling result varies depending on the criteria set. If we look at the state of respiratory arousal only when it is maintained for more than 3 seconds, there are cases where it is labeled as normal breathing even if 1 to 2 seconds of respiratory arousal are mixed. If the epoch is labeled as respiratory arousal even if it includes only 1 second, it is not easy to distinguish it from a normal breathing state because the respiratory arousal is only 1 second out of the total 30 seconds.

In order to solve the above problem, a new training dataset composition method is devised. First of all, we try to determine the appropriate clip length for the task. In the case of AHI, which is the criterion for determining OSA, it is evaluated how often apnea-hypopnea symptoms appear rather than how long the apnea-hypopnea was maintained. Therefore, it is important to detect short-lived respiratory arousal events well, since the short respiratory arousal event lasts only about 3 seconds. It is necessary to accurately classify that a respiratory arousal event has occurred even within 3 seconds. Therefore, dividing the clip into too short second-units increases the possibility that the respiratory arousal event is not included in one clip and is separated. However, if the length of the clip is too long, if the respiratory arousal occurs short, it may not be easy to distinguish because there is a difference of only a few seconds out of 30 seconds. Consequently, we constructed the dataset with a 30-second length, which is the basic unit of sleep research, and a 10-second length, which is utilized by benchmark datasets (Kinetics [15], Something-Something [9], UCF101 [22]), which are used for training most video action classification models. And what is a problem in the construction of sequential epoch units was the ambiguity of the labeling criteria. To solve this problem, the clip start time is determined based on the actual event occurrence time instead of dividing clips based on epoch time. For example, if the clip is 30 seconds long and the respiratory arousal occurs for 40 seconds, 30-second is extracted as a clip within that 40 seconds, and if it occurred for 15 seconds shorter than the clip length, a 30-second clip is created to include the corresponding 15-second event. In other words, the start time of the clip is also randomly designated so that the 15 seconds respiratory arousal events in the 30 seconds clip can be randomly arranged. That is, respiratory arousal event may appear from the beginning of the video, may appear during the middle 15 seconds, or may appear during the last 15 seconds. And, if respiratory arousal does not occur, 30 seconds are randomly sampled to create a clip.

Another thing to consider when creating a clip from a single video is that the frames in the video do not change significantly over time due to the characteristic of the sleep video. Therefore, when clips are extracted from consecutive time zones, the clips are almost similar and can be seen as duplicated data. To prevent this problem, a method of making only one clip per 10 minutes of video was taken. To summarize how clips and their labels are created, if there is a respiratory arousal event in a 10-minute video that is a candidate for clip generation, one 30-second clip with a respiratory arousal label is created in the 10-minute video. If no respiratory arousal event occurred during 10 minutes, a clip was created by randomly sampling 30 seconds between 1 minute and 10 minutes from the start of the video. The reason why only the last 9 minutes are used, excluding the previous 1 minute, is that there is a possibility that the respiratory arousal event that started at the end of the previous 10-minute video may be partially mixed in the first minute. When the clips are created in this way, the clips with the respiratory arousal labeling include only respiratory arousal events, and the others do not contain any respiratory arousal event.

Video Preprocessing In order to obtain high performance and fast inference time, video data should be preprocessed appropriately for the model without simply

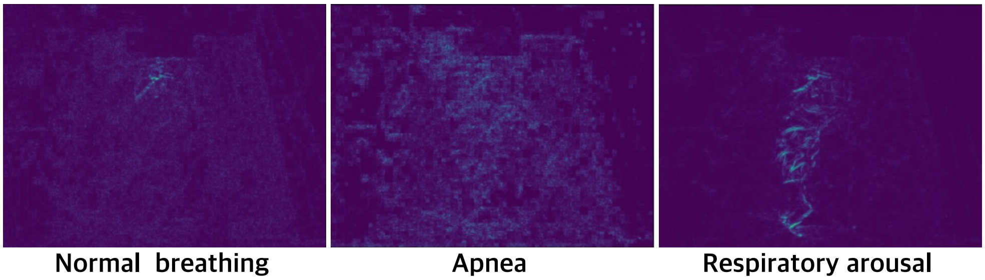
Normal breathing
Apnea
Respiratory arousal

Figure 3.4 Difference between frames by respiratory status

inserting the original video into the model. An important factor in the action recognition is to catch the changes from successive frame to frame. As a means to reflect such factors, in many cases, optical flow representing the movement pattern of objects is calculated and used as an input [20, 8, 6]. However, the optical flow is not suitable for the purpose of creating a lightweight detection system due to its high computational cost, and the movement of objects in sleep videos is monotonous and repetitive, so it does not require optical flow. Nevertheless, since observing the change between frames itself is an important factor in video analysis, it is possible to simply calculate the pixelwise value difference between frames and use it as an input. If the movement is big, such as respiratory arousal, the difference between the previous frame and the next frame will be large, and if the movement is small, such as apnea, there will be tiny change even if the frame is changed. Looking at Figure 3.4, in the case of apnea, there is little difference between frames, and it can be seen that the difference becomes clear in the order of normal breathing and respiratory arousal. One point to consider when calculating the difference between frames is how to compare the current frame with the frame after a certain point in time. It can be compared with the frame after 0.2 seconds or the difference with the frame after 2 seconds can be seen. In the respiratory arousal stage, the abdominal movement cycle is short because breathing is rapid. In normal breathing, the abdominal movement is relatively long. If too long a period is selected, not only respiratory arousal but also movement of normal breathing may be clearly captured and cannot be distinguished. If a short period is selected, respiratory arousal may also be indistinguishable because the difference between frames is small.

As mentioned earlier, due to the characteristic of sleep video, the frames in the video do not change significantly. In order to prevent overfitting of the training data due to the large amount of similar training data, the images were randomly flipped vertically and horizontally in each video clip.

For fast inference speed, it is necessary to reduce the size of input data while maintaining performance. At this time, factors that can reduce the size of the input clips include resolution and fps. When examining the differences between frames, appropriate resizing and frame skipping will be necessary as only the minimum resolution and frame rate needed to clearly see the differences. Due to the characteristics of sleep video, there is no significant change between frames, so there is little change in performance even when only a part of frames is used, and inference is faster when fewer frames are used. Therefore, only half of frames (2.5 fps) is used and each frame is resized to 96×96 resolution.

Lightweight Action Classification Video Model When a 30-second clip is given as an input, a model is needed to determine whether the subject in the video is in a state of respiratory arousal or not. As the task is simple with binary classification, it will work well enough even if the model is not large. Therefore, in the action classification task, we decide to utilize the MoViNets-A0 [16] architecture, which is known to be light enough and has good performance. Since the model operates at only 2.71 GFLOPs and 173 MB of memory, it corresponds to the purpose of this study to create a lightweight OSA detecting system.

3.3 OSA Classification

So far, we have focused on distinguishing whether or not respiratory arousal occurs at the video clip level. However, the fundamental goal of this study is to detect OSA. When there is a 6-hour sleep video for testing, the video is analyzed in clip length units, but the final result must be a classification result for OSA or not. To this end, a process of collecting the distinguishing results and final classification step should be added. If respiratory arousal is determined in units of 30-second clips, the ratio of the number of respiratory arousal to the total sleep time will be obtained. If a linear regression model is created by calculating the relationship between the ratio and AHI in the training data, the AHI can be estimated using the linear model in the test data as well. Then, using the estimated AHI values, whether the test subjects are OSA patients is detecetd.

4 Results

4.1 Detection of Respiratory Arousal within Video Clips

Event type	**Input data type**	**AUC**	**F1-score**
Respiratory Arousal	Frame difference	**0.825**	**0.80**
Respiratory Arousal	Original	0.523	0.72
Apnea-Hypopnea	Frame difference	0.558	0.74
Apnea-Hypopnea	Original	0.496	0.72

Table 4.1 Performance difference between event and input data type

Table 4.1 is a table comparing the performance of the Hospital A test set when the event detected by the model is respiratory arousal and when sleep apnea-hypopnea is used. Also, we can check the performance when the input data is used as the original and when the pixelwise value difference between frames is used. As can be seen from the table, the performance is the best with AUC of 0.825 and f1-score of 0.80, when the input data is used as the frame difference while detecting the respiratory arousal. Overall, the performance is better when the positive class is set as respiratory arousal and when the input data is also included as a difference between frames.

Looking at the performance when the original video is input as it is, it can be seen that the performance is very low whether it is detecting respiratory arousal or apnea-hypopnea. Through this, it can be confirmed that simply using the sleeping video as it is cannot be distinguished through the vision-based deep learning model. However, the performance improved when the frame difference was input. This is the point where we can see that it is necessary to observe any changes as the frames flow in order to clearly check whether an event has occurred in the sleep video. In the case of the apnea-hypopnea event, the performance improve by only 12.5% based on AUC even when the frame difference is observed, whereas in the case of the respiratory arousal event, the AUC increase by 57.74%. That is, it is confirmed through the results that the presence or absence of respiratory arousal show a more visually distinct difference than the presence or absence of apnea-hyopnea.

Clip unit	**Event type**	**AUC**	**F1-score**
Event	Respiratory Arousal	**0.825**	**0.80**
Epoch	Respiratory Arousal	0.568	0.32
Epoch	Apnea-Hypopnea	0.537	0.32

Table 4.2 Performance difference between clip unit and event type

Table 4.2 is a table comparing the performance when tested with successive 30-second epochs, the basic unit of sleep research, and the performance when tested with our own created clip method. As can be seen in the table, when data is input in units of epochs, the AUC is 0.568 and the f1-score is 0.32, resulting in poor performance. In other words, it can be seen that it is more effective to reconstruct the dataset by event in the case of a study such as this topic in which the occurrence of a specific event must be found, rather than a sleep study in which the continuous time flow is important, such as sleep stage classification.

Clip length	**AUC**	**F1-score**
10-second	0.723	0.75
30-second	**0.825**	**0.80**

Table 4.3 Performance difference according to clip length

Table 4.3 shows the performance difference when the clip length is 10 seconds and 30 seconds. The 10-second is a length commonly used in video action classification benchmark datasets, and the 30-second is a length commonly used in sleep studies. The performance when constructing the dataset for 30 seconds had an AUC of 0.825 and an f1-score of 0.8, which was higher than the performance at 10 seconds, AUC of 0.723 and an f1-score of 0.75. Due to the characteristic of sleep research, it is difficult for many changes to occur within a short time of 10 seconds, so the performance seems to be low because information in the clips is limited.

4.2 OSA Classification

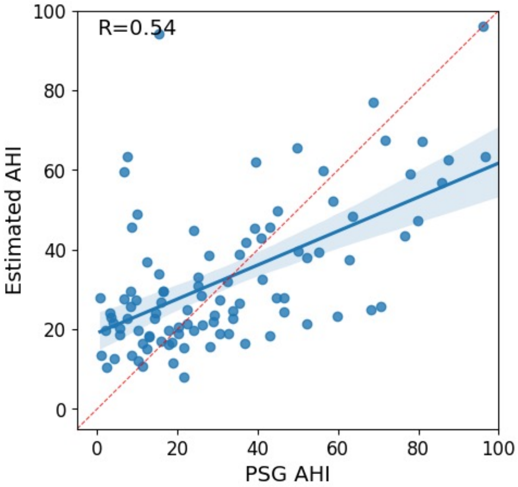

Figure 4.1 Scatterplots of PSG AHI vs estimated AHI values. The blue and red lines indicate fitted and unity lines, respectively

The results of estimating the AHI of the test case using the model that detects respiratory arousal can be seen in Figure 4.1. The Spearman correlation between the actual PSG AHI and the predicted AHI was calculated to be 0.54. Looking at the scatter plot, it is understood that the AHI is underestimated as a whole. This means that there are a large number of clips in which respiratory arousal actually occurred, but the model predicted that it did not. In other words, it seems that the error occurred because the performance of the respiratory arousal detection model itself is not very high. Conversely, in cases where the actual AHI is lower than 20, it is often predicted with a higher AHI than the actual one. In these cases, it seems that the AHI was estimated high because there were many clips that incorrectly predicted respiratory arousal even in non-respiratory arousal situations.

Method	Total cases	Accuracy	Precision	Recall	F1-score
Baseline [2]	41	82.93	77.78	95.45	0.86
Ours	499	74.00	73.91	97.14	0.84

Table 4.4 Performance of models on screening subjects with OSA

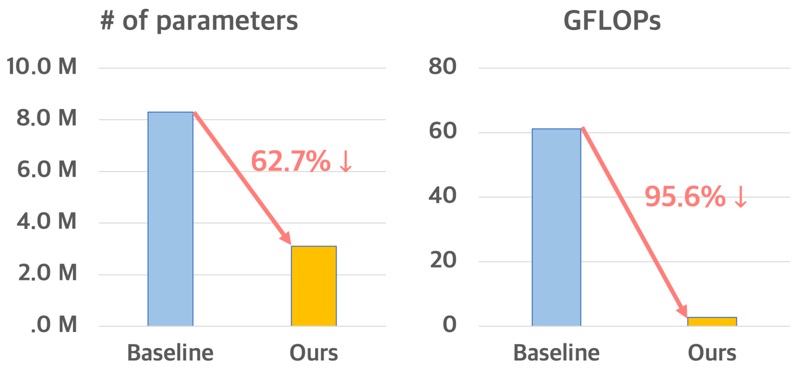

Figure 4.2 Comparison of model size between baseline and ours

In addition, the final OSA classification performance can be confirmed through Table 4.4, and the f1-score was 0.80 with an accuracy of 74.00%. This is lower than the baseline's 82.93% accuracy and the f1-score of 0.85. However, in the case of the baseline, there is a problem that it is difficult to see it as a generalized performance because the total number of cases used for training and testing is only 41 cases. In our research, the total number of cases is 499, 12 times more than the baseline, so our model can be considered a better generalized model. Also, if we look at the sleep video data, in the case of the baseline study, there is a problem that the reality is low because people are sleeping without covering the blanket and the camera is located right above the body. In other words, the experiment was conducted in a more predictable setting rather than the actual situation. If we reflect such differences, we can evaluate that although the accuracy is relatively low, we have achieved more realistically meaningful results. Furthermore, as shown in Figure 4.2, our model has a much smaller model size than the baseline model. The number of parameters is 62.7% less, and even 95.6% less for GFLOPs. Our model is more practical because it is small enough to run on edge devices.

4.3 Comparison Results on Other Hospital Data

Test data	**Data size**	**Video clips**		**OSA classification**			
		AUC	F1	Acc.	Precision	Recall	F1
Hospital A	100	0.825	0.80	**74.00**	73.91	97.14	0.84
Hospital D	118	0.731	0.82	**73.73**	96.05	72.28	0.82

Table 4.5 Comparison performance on Hospital A and Hospital D

All experiments so far have been the results of using only the case of Hospital A. Of the total 499 data of Hospital A, 399 cases were used for training and 50 cases were used for validation and test each. In the entire dataset, Hospital D was not used for training and was set as a test dataset to verify the generalization performance of the model. If the hospital that collects data changes, the environment for recording videos, measuring equipment, and reading technicians all change. Therefore, if the model is overfitting to Hospital A data, the performance will be lowered when experimenting with Hospital D data. In order to verify this generalization performance, training was performed with 399 cases of Hospital A and performance was verified through 118 cases of Hospital D, and the results are listed in Table 4.5. As can be seen from the table, there is no significant difference in performance even when the data source is changed. In other words, it can be seen that our model is well generalized.

5 Conclusion

We present a non-contact, vision-based obstructive sleep apnea detection method to overcome the limitations of current sleep apnea diagnosis. Existing OSA detection studies still has limitations such as having to sleep with the sensor attached or taking a long time for detection, but our system has the advantage of being able to obtain results quickly and conveniently regardless of location. In addition, for the first time, an attempt is made to utilize the presence or absence of respiratory arousal to detect OSA based on vision model. Furthermore, away from the consecutive epoch unit, which is the basic unit of sleep research, a new dataset is constructed in event units and used for training. In the preprocessing stage, the model is made lighter by reducing the computation cost by utilizing pixelwise value differences between frames rather than optical flow.

However, there is a disadvantage that accuracy is low compared to other methods. There are some cases where the AHI is overestimated or underestimated, and narrowing this difference will increase performance. In order to reduce the case of overestimation, other types of arousal that are easily misunderstood as respiratory arousal should be well detected. In order to reduce the case of underestimation, the accuracy of the model that detects respiratory arousal per clip should be improved. For this purpose, it would be a good attempt to devise a model structure suitable for the input data. In addition, from a data point of view, it seems good to configure the training data more densely as there were many missing data when constructing the training data.

Bibliography

[1] U. Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli: Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals. *Computers in Biology and Medicine*, **100** (2018), 270–278. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2017.09.017. URL: https://www.sciencedirect.com/science/article/pii/S0010482517303153.

[2] Sina Akbarian, Nasim Montazeri Ghahjaverestan, Azadeh Yadollahi, and Babak Taati: Noncontact sleep monitoring with infrared video data to estimate sleep apnea severity and distinguish between positional and nonpositional sleep apnea: model development and experimental validation. en. *J. Med. Internet Res.*, **23** (Nov. 2021), e26524.

[3] Md. Riyasat Azim, Shah Ahsanul Haque, Md. Shahedul Amin, and Tahmid Latif: Analysis of eeg and emg signals for detection of sleep disordered breathing events. *International Conference on Electrical & Computer Engineering (ICECE 2010).* 2010, 646–649. DOI: 10.1109/ICELCE.2010.5700776.

[4] Nannapas Banluesombatkul, Thanawin Rakthanmanon, and Theerawit Wilaiprasitporn: Single channel ECG for obstructive sleep apnea severity detection using a deep learning approach. *TENCON 2018 - 2018 IEEE Region 10 Conference.* IEEE, 2018. DOI: 10.1109/tencon.2018.8650429. URL: https://doi.org/10.1109%2Ftencon.2018.8650429.

[5] Arnab Bhattacharjee, Suvasish Saha, Shaikh Anowarul Fattah, Wei-Ping Zhu, and M. Omair Ahmad: Sleep apnea detection based on rician modeling of feature variation in multiband eeg signal. *IEEE Journal of Biomedical and Health Informatics*, **23** (2019), 1066–1074. DOI: 10.1109/JBHI.2018.2845303.

[6] João Carreira, and Andrew Zisserman: Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017, 4724–4733. DOI: 10.1109/CVPR.2017.502.

[7] Siyi Cheng, Chao Wang, Keqiang Yue, Ruixue Li, Fanlin Shen, Wenjie Shuai, Wenjun Li, and Lili Dai: Automated sleep apnea detection in snoring signal using long short-term memory neural networks. *Biomedical Signal Processing and Control*, **71** (2022), 103238. ISSN: 1746-8094. DOI: `https://doi.org/10.1016/j.bspc.2021.103238`. URL: `https://www.sciencedirect.com/science/article/pii/S1746809421008351`.

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox: Flownet: learning optical flow with convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[9] Raghav Goyal et al.: *The "something something" video database for learning and evaluating visual common sense.* 2017. DOI: `10.48550/ARXIV.1706.04261`. URL: `https://arxiv.org/abs/1706.04261`.

[10] Maziar Hafezi, Nasim Montazeri, Shumit Saha, Kaiyin Zhu, Bojan Gavrilovic, Azadeh Yadollahi, and Babak Taati: Sleep apnea severity estimation from tracheal movements using a deep learning model. *IEEE Access*, **8** (2020), 22641–22649. DOI: `10.1109/ACCESS.2020.2969227`.

[11] Sondre Hamnvik, Pierre Bernabé, and Sagar Sen: Yolo4apnea: real-time detection of obstructive sleep apnea. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20.* Edited by Christian Bessiere. Demos. International Joint Conferences on Artificial Intelligence Organization, July 2020, 5234–5236. DOI: `10.24963/ijcai.2020/754`. URL: `\url{https://doi.org/10.24963/ijcai.2020/754}`.

[12] Su Hwan Hwang, Chung Min Han, Hee Nam Yoon, Da Woon Jung, Yu Jin Lee, Do-Un Jeong, and Kwang Suk Park: Polyvinylidene fluoride sensor-based method for unconstrained snoring detection. *Physiological Measurement*, **36** (2015), 1399. DOI: `10.1088/0967-3334/36/7/1399`. URL: `https://dx.doi.org/10.1088/0967-3334/36/7/1399`.

[13] *Infrared sleep video data for diagnosing sleep disorders.* `https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=638`. Accessed: 2022-12-14.

[14] Arlene John, Koushik Kumar Nundy, Barry Cardiff, and Deepu John: Somnnet: an spo2 based deep learning network for sleep apnea detection in smartwatches. *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2021, 1961–1964. DOI: `10.1109/EMBC46164.2021.9631037`.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman: *The kinetics human action video dataset.* 2017. DOI: 10.48550/ARXIV.1705.06950. URL: https://arxiv.org/abs/1705.06950.

[16] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong: *Movinets: mobile video networks for efficient video recognition.* 2021. DOI: 10.48550/ARXIV.2103.11511. URL: https://arxiv.org/abs/2103.11511.

[17] Tanvir Mahmud, Ishtiaque Ahmed Khan, Talha Ibn Mahmud, Shaikh Anowarul Fattah, Wei-Ping Zhu, and M. Omair Ahmad: Sleep apnea detection from variational mode decomposed eeg signal using a hybrid cnn-bilstm. *IEEE Access*, **9** (2021), 102355–102367. DOI: 10.1109/ACCESS.2021.3097090.

[18] Mohammad Karimi Moridani, Mahdyar Heydar, and Seyed Sina Jabbari Behnam: A reliable algorithm based on combination of emg, ecg and eeg signals for sleep apnea detection : (a reliable algorithm for sleep apnea detection). *2019 5th Conference on Knowledge Based Engineering and Innovation (KBEI).* 2019, 256–262. DOI: 10.1109/KBEI.2019.8734992.

[19] Manish Sharma, Divyash Kumbhani, Anuj Yadav, and U Rajendra Acharya: Automated sleep apnea detection using optimal duration-frequency concentrated wavelet-based features of pulse oximetry signals. *Applied Intelligence*, **52** (Jan. 2022). DOI: 10.1007/s10489-021-02422-2.

[20] Karen Simonyan, and Andrew Zisserman: Two-stream convolutional networks for action recognition in videos. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1.* NIPS'14. Montreal, Canada: MIT Press, 2014, 568–576.

[21] Changyue Song, Kaibo Liu, Xi Zhang, Lili Chen, and Xiaochen Xian: An obstructive sleep apnea detection approach using a discriminative hidden markov model from ECG signals. en. *IEEE Trans. Biomed. Eng.*, **63** (July 2016), 1532–1542.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah: *Ucf101: a dataset of 101 human actions classes from videos in the wild.* 2012. DOI: 10.48550/ARXIV.1212.0402. URL: https://arxiv.org/abs/1212.0402.

[23] Jun-Sang Sunwoo, Young Hwangbo, Won-Joo Kim, Min Kyung Chu, Chang-Ho Yun, and Kwang Ik Yang: Prevalence, sleep characteristics, and comorbidities in a population at high risk for obstructive sleep apnea: a nationwide questionnaire study in south korea. en. *PLoS One*, **13** (Feb. 2018), e0193549.

[24] T.H. Tran, T.T. Nguyen, Z.M. Yuldashev, E.V. Sadykova, and M.T. Nguyen: The method of smart monitoring and detection of sleep apnea of the patient out of the medical institution. *Procedia Computer Science*, **150** (2019). Proceedings of the 13th International Symposium "Intelligent Systems 2018" (INTELS'18), 22-24 October, 2018, St. Petersburg, Russia, 397–402. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2019.02.069. URL: https://www.sciencedirect.com/science/article/pii/S1877050919304156.

[25] V Vimala, K Ramar, and M Ettappan: An intelligent sleep apnea classification system based on eeg signals. en. *J. Med. Syst.*, **43** (Jan. 2019), 36.

[26] Ching-Wei Wang, Andrew Hunter, Neil Gravill, and Simon Matusiewicz: Unconstrained video monitoring of breathing behavior and application to diagnosis of sleep apnea. en. *IEEE Trans. Biomed. Eng.*, **61** (Feb. 2014), 396–404.

[27] Yun Yin, Yule Hu, and Peizhi Liu: The research on denoising using wavelet transform. *2011 International Conference on Multimedia Technology*. 2011, 5177–5180. DOI: 10.1109/ICMT.2011.6002276.

[28] Junming Zhang, Zhen Tang, Jinfeng Gao, Li Lin, Zhiliang Liu, Haitao Wu, Fang Liu, and Ruxian Yao: Automatic detection of obstructive sleep apnea events using a deep CNN-LSTM model. en. *Comput. Intell. Neurosci.*, **2021** (Mar. 2021), 5594733.

[29] Kaiyin Zhu, Azadeh Yadollahi, and Babak Taati: Non-contact apnea-hypopnea index estimation using near infrared video. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Berlin, Germany: IEEE, July 2019.

초 록

우리는 빠르고 편하게 결과를 얻을 수 있는 비접촉 비전 기반 폐쇄성 수면 무호흡증 탐지 방법을 제시한다. 이를 위하여 세 가지 시도를 하는데, 첫째로 수면 연구의 기본 분석 단위인 에폭 중심에서 벗어나 이벤트 중심으로 새롭게 데이터셋을 구성한다. 둘째로, 수면 무호흡증 여부를 비전 기반으로 진단하기 위하여 호흡 각성 유무를 활용해보는 시도를 한다. 마지막으로 모델의 연산량을 줄이기 위하여 옵티컬 플로우를 활용하지 않고 프레임 간 픽셀 단위 값 차이를 활용한다. 그 뿐 아니라 기존 연구에서 사용된 데이터보다 몇 배 이상으로 많은 617건의 데이터를 활용하여 강건한 모델을 만들었다. 그 결과로 기준 모델에 비하여 플롭스 수가 95% 적으면서도 74% 정확도, 0.84의 f1 스코어를 달성했다.

주요어: 폐쇄성 수면 무호흡증, 호흡 각성, 무저호흡 지수, 비디오 분류

학 번: 2021-24878