electronics

MDPI

Article

Multi-Modality Tensor Fusion Based Human Fatigue Detection

Jongwoo Ha, Joonhyuck Ryu and Joonghoon Ko *

VAIV Company Inc., Seoul 04419, Republic of Korea; jwha@vaiv.kr (J.H.); jhryu@vaiv.kr (J.R.)
* Correspondence: joonghoon.ko@vaiv.kr

Abstract: Multimodal learning is an expanding research area and aims to pursue a better understanding of given data by regarding different modals. Multimodal approaches for qualitative data are used for the quantitative proofing of ground-truth datasets and discovering unexpected phenomena. In this paper, we investigate the effect of multimodal learning schemes of quantitative data to assess its qualitative state. We try to interpret human fatigue levels through analyzing video, thermal image and voice data together. The experiment showed that the multimodal approach using three types of data was more effective than the method of using each dataset individually. As a result, we identified the possibility of predicting human fatigue states.

Keywords: multimodal learning; fatigue detection; video analysis; tensor fusion; machine learning; human state; human detection

Citation: Ha, J.; Ryu, J.; Ko, J. Multi-Modality Tensor Fusion-Based Human Fatigue Detection. *Electronics* **2023**, *12*, 3344. https://doi.org/10.3390/electronics12153344

Academic Editor: Francesco Beritelli

Received: 30 June 2023
Revised: 28 July 2023
Accepted: 2 August 2023
Published: 4 August 2023

1. Introduction

Our world consists of numerous dynamic events. These events are compiled in a particular situation in a human sense or a similar way, which is called a modal. The data of an event can be observed in various modals, and the comprehensive judgment of these modals adds qualitative factors that are difficult to express quantitatively, such as emotions and external conditions, and appear as a result of the event. With the rise of multimodal AI [1], several studies have been conducted to analyze these modality data. In general, these research studies aimed to provide a better understanding of a given dataset from multiple modals. These works mainly intended to achieve specific purposes by learning multiple modality data together. In this study, we deal with thermal images and general video and voice data. Recently, there are many studies using multimodal fusion in the field of image analysis, which is covered in this study. Video data is often used in combination with text data. In general, methods for generating images from text or deriving text from images by applying a multimodal approach have been proposed [2]. There are also many other studies, for example, including image prediction using text data [3–7] and text prediction using image data [8]. In the case of thermal images, the multimodal learning method is mainly used for object or motion detection [9]. Voice data is also one of the most frequently used types of data in the multimodal approach, and is often used together with text data. In many cases, it is used for judgment problems about human emotions or facial expressions [10–12]. Furthermore, the research area is expanding to human state analysis [13–15] and language processing [16–18]. We intend to identify the possibility of analyzing the human fatigue state through the multimodal approach. The human fatigue state is qualitative, and it is very difficult to determine it via a single modality such as only recording thermal images or video or voice. We applied two approaches to the analysis of human fatigue using three modalities: video, thermal image and audio. In this study, we propose a method for classifying specific states of events expressed as qualitative data through the analysis of quantitative data and verifying it through experiments. We define video, voice, and thermal image data as different single modalities, and analyze them integrally adapting a tensor fusion network (TFN) to classify the fatigue level of a person, which is complex qualitative data. Our experiments show that

multimodal models outperform single-modal models. This means that analysis methods using multimodal data can extract (or imitate) qualitative judgments more effectively than those using single-modal models on the same data. In order to review this phenomenon, we separated the data input stage to compare and analyze the effect of each modality and its combination, and our experiments show that the higher the dimension of the multimodal model the higher the analysis performance. The outcomes that can be obtained through this study are as follows.

1. Through experiments, we explore whether a given problem can be viewed and solved from a complex perspective by combining multiple effects from data on the single-modal side. In particular, quantitative analysis results were intended to confirm whether they are effective in achieving indicators on qualitative evaluation criteria, providing implications for AI research such as empirical and intellectual judgment using complex human senses as inputs.
2. By comparing the effect of combining the results and analyzing each input data with a high-performance model to the effect of analyzing the representations of the raw input data, we propose an initial methodological study to discover new analysis orientations such as viewpoint and intention.

The remainder of this paper is organized as follows. In Section 2, we first address the related research studies concerning multimodal learning and measuring human fatigue. Section 3 defines each modality and its fusion method. Subsequently, the experimental results are described in Section 4. Section 5 provides a brief overview of this study and the experiments.

2. Related Works

2.1. Multimodal Learning

Multimodal learning or multimodal AI is an expanding research area. There are many study areas, including image tagging [3], image captioning [4,5], text-to-image [6,7], visual question answering [8], scene recognition by sound [11], and Speech2Face [12]. A common aim of these various studies is to improve the effectiveness of existing modals through the addition of other modals, or to enable the analysis of existing modals to reach new targets.

Converging multiple modals can be said to be a basic approach to problem solving through multimodal learning. The most basic method is to connect feature vectors derived from each modal, which is called early fusion [19–22].

However, multimodal model learning is a challenging task due to differences in format or representation among the modalities, such as differences in feature space, and can result in biased results for specific modalities. Recent attempts at multimodal learning have suggested the importance of the proper use of tensors for multimodal representations [23]. Among the studies, we applied the TFNs [24] method to represent and control the multi-modals through the meaningful manipulation of tensors to propose adaptive models.

2.2. Measuring Fatigue and Fatigue Levels

Fatigue can occur due to various causes such as human physiological characteristics, sleep disorders, lifestyle habits, and stress. The most important factor is sleep disturbance, which can be assumed to be fatigue in situations where performance is degraded because sleep time affects concentration and decision-making [25].

Various methods such as questionnaires, biochemical/physiological tests, and response tests are used to measure biological fatigue levels. For the questionnaire, statistically significant results were shown when developing fatigue measurement tools based on the multi-dimensional fatigue inventory (MFI) [26,27]. Actigraphy devices, electrocardiography (ECG), and ecological instantaneous evaluation using skin temperature have also been used [28,29].

These methods can include changes in levels due to differences in subjective judgment and personal situations, and rarely set a categorical level of fatigue. Moreover, measuring

through surveys and questionnaires requires a lot of resources, such as time, and cannot ignore the effects of external factors such as nutrition, exercise, and infection [30].

Some research works are ongoing concerning data collection and analysis related to fatigue, such as the driver drowsiness detection dataset [31]. However, these studies mainly focus on unimodal (single modality) motion detection, not multimodal approaches or fatigue detection. We propose a fatigue level measurement method using a multimodal approach to verify the efficiency of each known modality data.

2.3. Findings on Human Activities

One meaning in a situation can appear as an action.Motion detection in an image constituting the video is used to find its meaning or emotion. HOG (Histogram of Gradient) [32] uses images of faces shot from the front as well as images of natural poses. The UCF101 dataset [33] defines 101 activity classes and contains 13,000 clips and 27 h of image data, which are used to analyze human activity. Recent studies have used Convolution-3D [34]. We propose quantitative analysis methods to find comprehensive meanings such as fatigue, rather than practical generalized methods such as behavior/activity recognition.

3. Detecting Fatigue Levels through Multimodal Tensor Fusion

We propose a fatigue level measurement model (F-TFN) that works through the tensor fusion method. In order to derive fatigue levels, which are data expressed in qualitative states, multimodal learning methods using three modalities, namely image, voice, and thermal images, expressed in quantitative data. To train this model, we tried to derive fatigue levels by appropriately fusing features or processing three input modalities to tensors via SubModel or SubNet for each input data.

Our method adopted a TFN tensor fusion method, however, there are differences in the type of input data and feature extraction or data processing method for each input type. The input tensor of TFN is a feature derived by applying a new model (SubModel) for each modality and vector embedding. However, in this work, we focused on simply reducing the dimensions of the input data and representing the original data as it is, rather than extracting the feature for each modality by applying a separate SubNet. These methods will be meaningful in solving real-world problems, including qualitative data that are difficult to express with tensors.

Therefore, we define our first model (F-TFN1) by defining the input process proposed by the base TFN as a SubModel method, and we propose a second model (F-TFN2) by applying and utilizing the newly proposed SubNet method in this study. We performed comparative experiments on these two models. In addition, each system was compared and analyzed through experiments in the unimodal scheme, deriving the fatigue levels using a single modality. In the unimodal scheme, the fatigue level was derived through the FCL (Fully Connected Layers) of the input tensor of modality to determine whether each modality could derive a fatigue level separately without interference from the other modalities. In the existing TFN approach, when the method of pre-processing and fusion of each modality input data is applied to our high-dimensional data, the loss of original data information is inevitable. Therefore, we propose a second method (F-TFN2) that applies only the data reduction method for fusion to each input in order to maintain the information of the original data as much as possible.

First, the input data of each modal is input into the corresponding SubModel (or SubNet) and converted into a processed tensor. Each tensor then expands the dimension, filling it with values of 1 for the TFN operations. Inside the TFN model, we generate a new combined tensor, z_{fusion}, through operations on three inputs using the Cartesian product method. When expanding the dimension of the tensor for each modality, we fill the dimension with values of 1 so as we can preserve both unimodal and biomodal tensors. Therefore, the fused tensor retains the existing information corresponding to the unimodal and biomodal tensors as well as the newly generated trimodal tensor by combining the

three inputs. The fused tensor is then input into the FCL and finally a single sigmoid out layer is applied to derive the final fatigue level.

3.1. F-TFN1: SubModels

Each modal represented by quantitative data has its own different forms and dimensions. To ensure that reasonable tensor fusion is achieved in this situation, we first apply a SubModel method, extracting features from the original data of each modal. The SubModel method is a method of applying a specific model to the input, extracting the appropriate feature, vector-embedding, and defining it as an input tensor. For three inputs, video, speech, and thermal images, the SubModel generates a feature tensor z_i for the fusion of the three modals. Figure 1 shows the tensor fusion process applied with the SubModel and the fusion tensor z_{fusion} newly expressed through the tensor fusion.

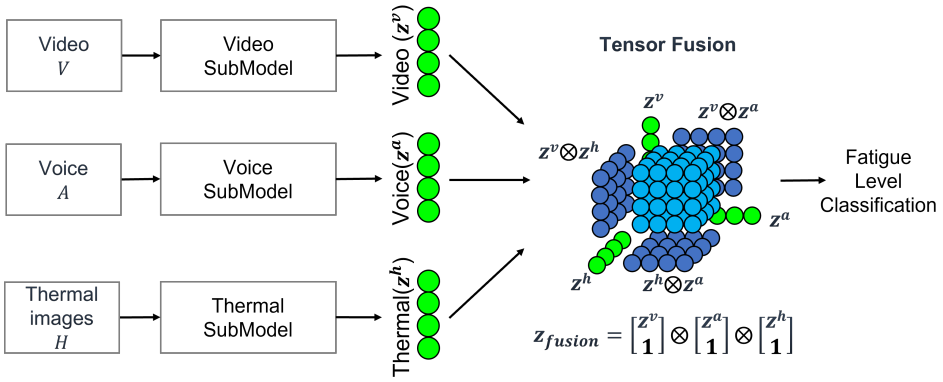

Figure 1. F-TFN1: SubModels for extracting modality features.

Video-Embedding SubModel: the video data were taken by a typical camera from the front with the target sitting for one minute, with an original resolution of 1280 × 720. During the analysis process, one video datum was decomposed into k frames, which was reduced to a three-channel image of $(H, W, 3)$ size for analysis. To extract the feature of the image, the SubModel applied EfficientNet, showing the highest performance in the image analysis among the CNN series. Feature z_i for the entire video was expressed by extracting the features of all the decomposed frames and connecting them. Figure 2 shows the feature extraction process for each frame f_i^j in the i-th video data. The EfficientNet-B0 model was applied in the feature extraction process. Equation (1) shows the above process.

$$\begin{aligned} & z_i = (z_i^1, z_i^2, \ldots, z_i^k), \quad |z_i| = k * |z_i^j| \\ Where \quad & 0 < i \leq n \quad (n : Dataset\ size) \\ & 0 < j \leq k \end{aligned} \tag{1}$$

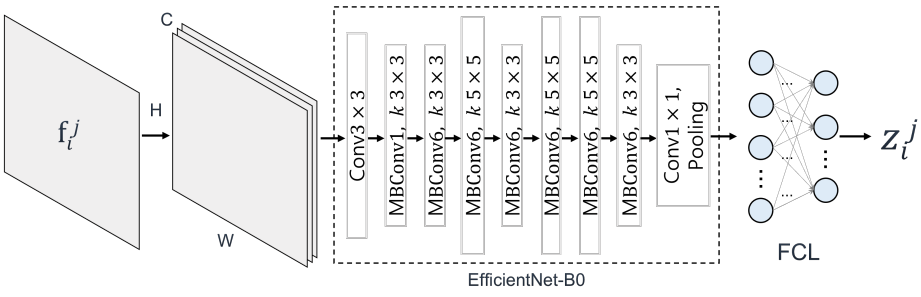

Figure 2. Video SubModel.

Voice-Embedding SubModel: first, we extract the voice data from the video data. Furthermore, the one minute raw voice data are represented by a vector size of T = 220,000.

In the case of voice data, the reduction ratio of the voice data is most important for tensor fusion with the available resources because its size is very large compared to the video and thermal image data. In the case of using SubModel for the voice data, to extract features we proposed a model that extracts the F0-mean value, a representative indicator of the voice data. As shown in Figure 3, we extracted F0-mean values per specific interval t for the i-th voice datum a_i, combining them to generate a feature tensor z_i of the corresponding voice data. The combination process of the extracted F0-mean values is the same as in Equation (2).

$$z_i = (F0_i^1, F0_i^2, \ldots, F0_i^{\frac{T}{t}}) \quad Where \quad 0 < i \leq n \quad (n : Dataset\ size) \tag{2}$$

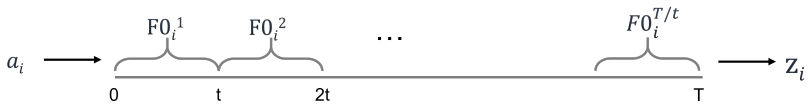

Figure 3. Voice SubModel.

Thermal Image-Embedding SubModel: the thermal data are images taken using a thermal camera for one minute. An average of 300 thermal images were taken per minute, and in this study, k thermal images, the same as the number of frames in the video data, were randomly chosen. In the case of thermal images, they have the same data form as a frame of the video data, so the SubModel of the thermal data was the same as the method for the video data, generating an input tensor z_i of the thermal data.

3.2. F-TFN2: SubNets

We proposed F-TFN1, a tensor fusion method that works through feature extraction using SubModel. For F-TFN1, we extract features for each input, but for the newly proposed F-TFN2 model, the goal is to define an input tensor for fusion by using SubNet to reduce and process the data while keeping the meaning of the input data. There is a big difference in the shape and size of each piece of input image, voice, or thermal data. Therefore, in this method, the problem of input data inequality should also be considered. We propose a model, F-TFN2, that uses these SubNet and carry out a comparative experiment using the F-TFN1 model. Figure 4 shows the tensor fusion process of the F-TFN2 to which SubNet is applied.

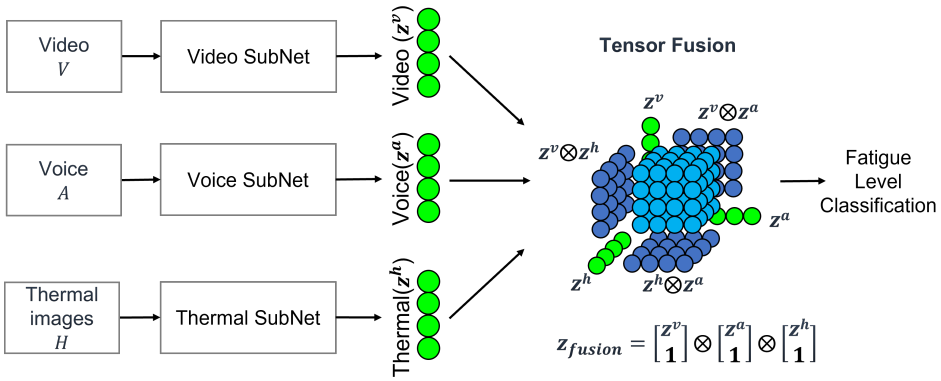

Figure 4. F-TFN2: SubNets for modality embedding.

Video-Embedding SubNet: Figure 5 shows a SubNet for video data processing. In this case, k frames are decomposed from the input video. Furthermore, each frame's channel is expanded through DenseNet. Next, simple operations using a global average pooling

(GAP) layer reduce the frames to a one-dimensional vector z_i^j. The generation process of z_i is the same as in Equation (3).

$$\begin{aligned} & z_i = (z_i^1, z_i^2, \ldots, z_i^k), \quad |z_i| = k * |z_i^j| \\ Where \quad & 0 < i \leq n \quad (n : Dataset\ size) \\ & 0 < j \leq k \end{aligned} \tag{3}$$

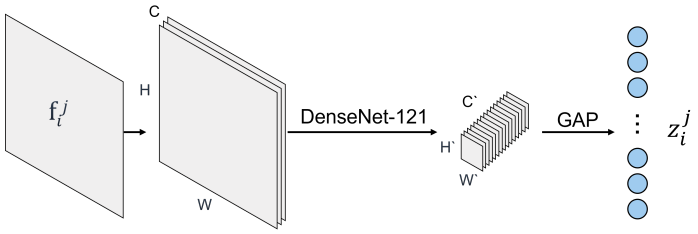

Figure 5. Video subnetwork.

Voice-Embedding SubNet: SubNet for voice data works with the same goal. SubNet for input data processing is similarly only applied to a simple form of the fully connected layer (FCL) to produce a processed input tensor that can shrink the input data to keep the original information. We applied an output layer of the FCL so that the size $|z_i|$ of the voice tensor can be generated with an equal size to the output tensor in the SubModel method.

Thermal Embedding SubNet: SubNet for processing the thermal image data also applied the same as the method of video data, generating an input tensor z_i of the thermal image data.

4. Results

4.1. Fatigue Level Dataset

The bio-signal collection system [35] is specifically developed to collect human signals associated with fatigue. For this study, the collection subjects were selected, and data relating to their fatigue status were collected. Figure 6 represents the data collection process.

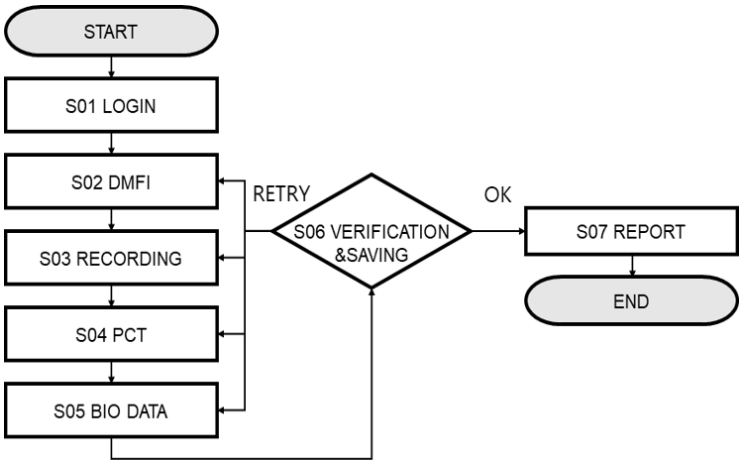

Figure 6. Data collection process.

In the "S03 RECORDING" stage in Figure 6, information for data collection is displayed in the window of the data collection device, such as in Figure 7. At this time, guidelines for data collection, such as the direction of the face and a script for voice recording, are provided to the subject through the screen. We simultaneously measured three types of data: video, thermal image, and audio for one person for one minute. That is, one fatigue level data includes three types of collected data. The subjects were asked to read a given script for about one minute, and video and thermal images were recorded after the

start. After recording, the subjects selected their subjective fatigue level from 1 (excellent) to 5 (extreme fatigue).

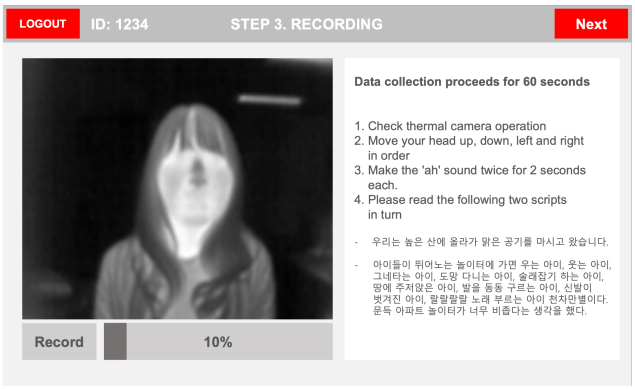

Figure 7. Display sample of the biological signal acquisition system.

Next, the researchers analyzed the actual fatigue using the daily multi-phasic fatigue inventory (DFMI), psychomotor cognitive test (PCT), and blood and saliva samples. Through this process, data were collected and used for experiments, and one dataset includes one minute of video, voice, thermal and fatigue levels (1 to 5). In the data collection process, a total of 7000 datasets were collected for 292 subjects. In this work, 690 datasets were used for each label from 1 to 5 according to the label with the smallest number of datasets to balance the data. Therefore, $n = 3450$ data points were used in the experiment.

4.2. Model Tuning

Our experiments are intended to classify fatigue levels from 1 to 5 and all models were trained using cross-entropy loss. Adam was applied as an optimizer, and learning was conducted for 100 epochs with a batch size of 16. The specific SubModel/SubNet configuration environment is as follows.

Video/Thermal-Embedding SubModel/SubNet: the video data were taken by a typical camera from the front with the data collection target sitting for one minute, with a resolution of 1280×720. The same method was applied because the thermal data also had a similar shape to the video data. In this study, the number of frames used per datum was set to $k = 60$, and each frame was reduced to three-channel images of $W = 224, H = 224$ according to the EfficientNet input layer. In the SubModel method, the output layer of the FCL was adjusted so that the size of the processed tensor z_i^j for each frame was 16. For the video SubNet, we applied the DenseNet-121 model to extend the channel of the frame to sizes of $W' = 7$, $H' = 7$, and $C' = 16$, and the size of the final output vector was determined via GAP.

Voice-Embedding SubModel/SubNet: for the voice data, the size of the original data was very large at T = 220,000. In this case, a very high reduction ratio was required compared to the other two modals to solve the resource limit problem in the fusion operation. In this study, for the SubModel, we created a feature by setting the extraction range of the F0-mean value to $t = 10,000$. In SubNet, we made the output size of the FCL (22) so that the fabricated tensor size of voice data was $|z_i| = 22$.

4.3. Experimental Results

In the fatigue level classification experiment, the performance in the unimodal and multimodal (TFN) environments according to the combination of the three modalities, videos, voices, and thermal images, are shown in Table 1. The evaluation of the models used a test dataset comprising 20% of the total data. The experimental results showed that the classification performance in the multimodal system increased significantly compared to that in each single-modal classification system. Human fatigue is a value of a qualitative state, and there is a very high probability that subjective judgment will intervene in the

data collection of the subject. From this perspective, it can be seen that the multimodal system can reflect human senses or emotions more realistically through the experimental results. In addition, the possibility of multimodal systems can be identified in that new qualitative states can be classified through quantitative state images, thermal images, and voice data.

Table 1. Performance evaluation based on the multimodal methodology.

Environment	Modalities	F-TFN 1 (SubModel)			F-TFN 2 (SubNet)		
		Accuracy	Recall	Precision	Accuracy	Recall	Precision
Uni-modality	Videos	0.359	0.185	0.332	0.401	0.167	0.498
	Voices	0.311	0.334	0.327	0.345	0.463	0.324
	Thermal images	0.320	0.241	0.263	0.317	0.280	0.370
Bi-modality	Videos + Voices	0.419	0.409	0.637	0.424	0.455	0.553
	Videos + Thermal images	0.452	0.382	0.570	0.381	0.315	0.453
	Voices + Thermal images	0.403	0.431	0.425	0.457	0.553	0.497
Tensor Fusion	Videos + Voices + Thermal images	0.598	0.703	0.651	0.646	0.742	0.717

5. Conclusions

In this paper, we studied the effectiveness of multimodal learning schemes on quantitative data to evaluate the qualitative state of fatigue. Specifically, these schemes were used to measure the current fatigue level of humans by analyzing video, voice, and thermal data together. Three types of data were defined for each modality, and three modalities were applied to the F-TFN model to generate a new input tensor. We trained the model to classify the fatigue levels using this newly generated tensor. The experiments show that fatigue level classification in multimodal systems outperforms uni-modality analysis systems. We experimentally demonstrated the effectiveness of multimodal data fusion. The performance of the current model shows an accuracy of about 65%, and it is necessary to improve the performance for realistic application. With these results, we identified the predictability of human conditions. We are trying to apply this fatigue measurement method to a situation where fatigue management is required. Our goal is to provide a human fatigue measurement platform with an accuracy of over 85%. In order to increase the utilization of the system, the first step is to increase the reliability of the fatigue measurement system, that is, the performance of the analysis model. In order to improve the performance of the fatigue analysis engine, follow-up research that applies advanced feature extraction techniques or minimizes the loss of original data in the process of data reduction must be conducted.

Author Contributions: Methodology, J.H.; Formal analysis, J.H. and J.R.; Data curation, J.R.; Writing—original draft, J.H.; Writing—review & editing, J.K.; Supervision, J.K.; Project administration, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Civil-Military Dual Use Technology Development Work (No. 20-CM-BD-13) of Institute of Civil Military Technology Cooperation (ICMTC), funded by ROK Ministry of Trade, Industry and Energy and Defense Acquisition Program Administration.

Data Availability Statement: The datasets generated and/or analyzed during the current study are not publicly available for military purposes and for privacy reasons.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Ramachandram, D.; Taylor, G.W. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process. Mag.* **2017**, *34*, 96–108. [CrossRef]
2. Seo, P.H.; Nagrani, A.; Arnab, A.; Schmid, C. End-to-end generative pretraining for multimodal video captioning. In Proceedings of the of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–22 June 2022; pp. 17959–17968.
3. Fu, J.; Rui, Y. Advances in deep learning approaches for image tagging. *APSIPA Trans. Signal Inf. Process.* **2017**, *6*, E11. [CrossRef]
4. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3156–3164.
5. Mathur, P.; Gill, A.; Yadav, A.; Mishra, A.; Bansode, N.K. Camera2Caption: A real-time image caption generator. In Proceedings of the International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 2–3 June 2017; pp. 1–6.
6. Qiao, T.; Zhang, J.; Xu, D.; Tao, D. Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2020; pp. 1505–1514.
7. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8821–8831.
8. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.
9. Wu, D.; Cao, L.; Zhou, P.; Li, N.; Li, Y.; Wang, D. Infrared small-target detection based on radiation characteristics with a multimodal feature fusion network. *Remote Sens.* **2022**, *14*, 3570. [CrossRef]
10. Rana, A.; Jha, S. Emotion based hate speech detection using multimodal learning. *arXiv* **2022**, arXiv:2202.06218.
11. Aytar, Y.; Vondrick, C.; Torralba, A. Soundnet: Learning sound representations from unlabeled video. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
12. Oh, T.H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W.T.; Rubinstein, M.; Matusik, W. Speech2face: Learning the face behind a voice. In Proceedings of the of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7539–7548.
13. Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176.
14. Zadeh, A.; Zellers, R.; Pincus, E.; Morency, L.P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intell. Syst.* **2016**, *31*, 82–88. [CrossRef]
15. Poria, S.; Cambria, E.; Gelbukh, A. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2539–2544.
16. Elliott, D.; Kiela, D.; Lazaridou, A. Multimodal learning and reasoning. In Proceedings of the of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Berlin, Germany, 6–12 August 2016.
17. Chen, C.; Han, D.; Wang, J. Multimodal encoder-decoder attention networks for visual question answering. *IEEE Access* **2020**, *8*, 35662–35671. [CrossRef]
18. Verma, G.; Mujumdar, R.; Wang, Z.J.; De Choudhury, M.; Kumar, S. Overcoming Language Disparity in Online Content Classification with Multimodal Learning. *arXiv* **2022**, arXiv:2205.09744.
19. Hou, J.C.; Wang, S.S.; Lai, Y.H.; Tsao, Y.; Chang, H.W.; Wang, H.M. Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Trans. Emerg. Top. Comput. Intell.* **2018**, *2*, 117–128. [CrossRef]
20. Rastgoo, M.N.; Nakisa, B.; Maire, F.; Rakotonirainy, A.; Chandran, V. Automatic driver stress level classification using multimodal deep learning. *Expert Syst. Appl.* **2019**, *138*, 112793. [CrossRef]
21. Ma, Y.; Hao, Y.; Chen, M.; Chen, J.; Lu, P.; Košir, A. Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Inf. Fusion* **2019**, *46*, 184–192. [CrossRef]
22. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.H.; Chang, S.F.; Cui, Y.; Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In Proceedings of the Advances in Neural Information Processing Systems, Virtual Event, 6–14 December 2021; Volume 34.
23. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.; Morency, L.P. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv* **2018**, arXiv:1806.00064.
24. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. *arXiv* **2017**, arXiv:1707.07250.
25. Dinges, D.F. An overview of sleepiness and accidents. *J. Sleep Res.* **1995**, *4*, 4–14. [CrossRef]
26. Smets, E.; Garssen, B.; Bonke, B.d.; De Haes, J. The Multidimensional Fatigue Inventory (MFI) psychometric qualities of an instrument to assess fatigue. *J. Psychosom. Res.* **1995**, *39*, 315–325. [CrossRef]
27. Lee, Y.; Shin, S.; Cho, T.; Yeom, H.; Kim, D. An Experimental study on Self-rated Fatigue Assessment Tool for the Fatigue Risk Groups. In Proceedings of the of 2021 KMIST Conference, Virtual Event, 11–15 January 2021; pp. 1755–1756.
28. Choe, J.H.; Antoine, B.S.R.; Kim, J.H. Trend of Convergence Technology between Healthcare and the IoT. *Inf. Commun. Mag.* **2014**, *31*, 10–16.
29. Kim, K.; Lim, C. Wearable health device technology in IoT era. *Korean Inst. Electr. Eng.* **2016**, *65*, 18–22.

30. Kim, D. A Study on the Pilot Fatigue Measurement Methods for Fatigue Risk Management. *Korean J. Aerosp. Environ. Med.* **2020**, *30*, 54–60. [CrossRef]
31. Weng, C.H.; Lai, Y.H.; Lai, S.H. Driver drowsiness detection via a hierarchical temporal deep belief network. In Proceedings of the Asian Conference on Computer Vision, Taipei, Taiwan, 20–24 November 2016; pp. 117–133.
32. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1; pp. 886–893.
33. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.0402.
34. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
35. Yoo, S.; Kim, S.; Kim, D.; Lee, Y. Development of Acquisition System for Biological Signals using Raspberry Pi. *J. Korea Inst. Inf. Commun. Eng.* **2021**, *25*, 1935–1941.