electronics

MDPI

Article

AffectiVR: A Database for Periocular Identification and Valence and Arousal Evaluation in Virtual Reality

Chaelin Seok [1], Yeongje Park [1], Junho Baek [1], Hyeji Lim [1], Jong-hyuk Roh [2], Youngsam Kim [2], Soohyung Kim [2] and Eui Chul Lee [3,*]

[1] Department of AI & Informatics, Graduate School, Sangmyung University, Seoul 03016, Republic of Korea; 202334018@sangmyung.kr (C.S.); 202331043@sangmyung.kr (Y.P.); 202232029@sangmyung.kr (J.B.); 202232034@sangmyung.kr (H.L.)
[2] Cyber Security Research Division Electronics and Telecommunications Research Institute (ETRI), Daejeon 34129, Republic of Korea; jhroh@etri.re.kr (J.-h.R.); kimzt@etri.re.kr (Y.K.); lifewsky@etri.re.kr (S.K.)
[3] Department of Human-Centered Artificial Intelligence, Sangmyung University, Seoul 03016, Republic of Korea
* Correspondence: eclee@smu.ac.kr

Abstract: This study introduces AffectiVR, a dataset designed for periocular biometric authentication and emotion evaluation in virtual reality (VR) environments. To maximize immersion in VR environments, interactions must be seamless and natural, with unobtrusive authentication and emotion recognition technologies playing a crucial role. This study proposes a method for user authentication by utilizing periocular images captured by a camera attached to a VR headset. Existing datasets have lacked periocular images acquired in VR environments, limiting their practical application. To address this, periocular images were collected from 100 participants using the HTC Vive Pro and Pupil Labs infrared cameras in a VR environment. Participants also watched seven emotion-inducing videos, and emotional evaluations for each video were conducted. The final dataset comprises 1988 monocular videos and corresponding self-assessment manikin (SAM) evaluations for each experimental video. This study also presents a baseline study to evaluate the performance of biometric authentication using the collected dataset. A deep learning model was used to analyze the performance of biometric authentication based on periocular data collected in a VR environment, confirming the potential for implicit and continuous authentication. The high-resolution periocular images collected in this study provide valuable data not only for user authentication but also for emotion evaluation research. The dataset developed in this study can be used to enhance user immersion in VR environments and as a foundational resource for advancing emotion recognition and authentication technologies in fields such as education, therapy, and entertainment. This dataset offers new research opportunities for non-invasive continuous authentication and emotion recognition in VR environments, and it is expected to significantly contribute to the future development of related technologies.

Keywords: biometric authentication; head-mounted display; infrared periocular images; periocular dataset

check for updates

Citation: Seok, C.; Park, Y.; Baek, J.; Lim, H.; Roh, J.-h.; Kim, Y.; Kim, S.; Lee, E.C. AffectiVR: A Database for Periocular Identification and Valence and Arousal Evaluation in Virtual Reality. *Electronics* **2024**, *13*, 4112. https://doi.org/10.3390/electronics13204112

Academic Editors: Zhaofeng He, Yunlong Wang, Caiyong Wang, Jianze Wei and Min Ren

Received: 25 September 2024
Revised: 15 October 2024
Accepted: 17 October 2024
Published: 18 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The global head-mounted display (HMD) market was valued at approximately USD 22.31 billion in 2022 and is expected to grow at a compound annual growth rate (CAGR) of 35.8% from 2023 to 2028 [1]. This growth indicates the accelerated adoption and technological advancements of HMDs across various fields such as virtual reality (VR), augmented reality (AR), gaming, medical applications, and military training. Immersion is a key element of VR experiences, playing a critical role in determining the quality of user experience. Higher immersion levels allow users to have more natural and realistic experiences in VR environments, significantly enhancing educational outcomes, therapeutic effects, and satisfaction in entertainment [2]. To increase immersion, various technical elements,

including field of view, resolution, user recognition, and feedback, are crucial. These elements enhance interaction and realism in virtual environments, enabling deeper immersion. In particular, in order to maximize immersion, interaction in a VR environment must be natural and uninterrupted, and for this, implicit continuous authentication and emotion recognition play an important role.

Implicit continuous authentication is a method that allows authentication to occur naturally without the user being aware of the process, maintaining security without disrupting immersion in VR environments [3]. This approach continuously verifies user identity during a session, enhancing security and convenience by removing the need for repetitive actions like entering personal identification numbers(PINs). Periocular biometrics, which utilizes features around the eyes—such as eyelids, eyelashes, eyebrows, lacrimal glands, eye shape, and skin texture—holds significant potential for this in HMDs [4]. The periocular region offers high recognition accuracy with simple data acquisition [5]. In HMDs, the eye area remains fixed, making the process resilient to motion noise and ensuring consistent authentication. Emotion evaluation plays an important role in enhancing the realism of VR content. It helps develop better VR environments and uses the emotions felt by users as key data for evaluating VR content [6]. Emotion recognition enables VR systems to interact with users more naturally, creating environments that foster deeper immersion. For example, in educational VR content, emotion recognition monitors users' understanding or interest in real-time, allowing dynamic adjustments to the learning content, thereby maximizing learning effectiveness [7].

However, the development of technologies to enhance immersion still faces significant barriers due to a lack of suitable data. Most existing public datasets do not include periocular data obtained from VR devices, primarily relying on data captured by external cameras. Notable datasets like the CASIA Iris Database, IIT Delhi Dataset, and ND-Iris 0405 contain data collected with external equipment, not cameras embedded in headsets used in VR environments [8–10]. Thus, these datasets fail to reflect the specific conditions of VR, limiting their applicability in VR-based user authentication and emotion evaluation research. To overcome these limitations, this study collected data from 100 participants wearing VR devices. The high-resolution data accurately capture the features of the periocular region, providing essential material for user authentication and related research in VR environments. Additionally, participants watched seven emotion-inducing videos, with their emotions recorded according to Russell's emotion model, creating a dataset that can be used for future emotion evaluation in VR. This study presents a baseline for periocular biometric using the collected data, demonstrating the potential for its use in various future research areas.

2. Related Works

2.1. Dataset

Because it is difficult to acquire data in a VR environment, the data acquired in a VR environment are limited. Therefore, we compared our dataset with some existing datasets, including a periocular dataset that was not acquired in a VR environment. A comparison is presented in Table 1.

High-quality images are required for accurate periocular authentication. However, most public datasets have a low resolution of 640 × 480 or less, and among public datasets, high-resolution datasets have a low number of subjects. In addition, because most datasets obtained from VR environments were collected in the initial state of wearing VR, there is a lack of consideration of actual use environments, such as situations where the user takes off the VR device and puts it back on. However, in this study, we proceeded with the process of wearing the VR device again each time we watched each video. Therefore, a dataset was constructed taking into account changes in the eye position in the image depending on the device location.

Table 1. Comparison of periocular datasets.

Dataset	#Images	#Participants	Resolution	Camera/Sensor	Environment
CASIA-IrisV4 [8]	54,607	2800	640 × 480	-	Non-VR
IIT Delhi Iris Database [9]	1120	224	320 × 240	JIRIS, JPC1000, digital CMOS camera	Non-VR
ND-Iris 0405 [10]	64,980	356	640 × 480	LG 2200 iris imaging system	Non-VR
UBIRIS.v1 [11]	1877	241	300 × 300	Nikon E5700	Non-VR
UBIRIS.v2 [12]	11,002	261	72 × 72	Canon EOS 5D	Non-VR
MRL Eye Dataset [13]	84,898	37	640 × 480	Intel RealSense RS 300 sensor	Non-VR
			1280 × 1024	IDS Imaging sensor	Non-VR
			752 × 480	Aptina sensor	Non-VR
OpenEDS [14]	356,649	152	400 × 640	-	VR
VISA Dataset [15]	3501	100	640 × 480	IriSheild Camera	Non-VR
NVGaze [16]	7400	30	640 × 480	-	VR
OpenEDS2020 [17]	550,400	90	640 × 400	-	VR
Our Dataset	5,199,175	100	1920 × 1080	HTC Vive Binocular Add-on	VR

2.2. User Authentication Using Head-Mounted Display

Liebers et al. [18] proposed a research direction that utilizes behavior-based eye biometrics to enhance security in VR environments as more and more HMDs with built-in eye tracking functions are released to the market. Eye tracking technology must accurately track and analyze the user's eye movements or patterns. However, for some users, there may be issues with the accuracy and reliability of the authentication system because not all users respond comfortably to eye tracking technology. Luo et al. [19] pointed out that more and more personal and sensitive data are being generated due to the spread of VR and proposed a new biometric authentication method that adapts the human visual system (HVS) to the VR platform. OcuLock [20] is a system that combines an electro-oculography (EOG)-based HVS detection framework and a record comparison-based authentication scheme. This system considers the entirety of the HVS, including the eyelids, eye muscles, cells, and peripheral nerves, as well as eye movements. OcuLock achieved low equivalent error rates (EERs) of 3.55% and 4.97%. However, the OcuLock system uses EOG-based sensors to detect the HVS. Therefore, it has the disadvantage of requiring a separate sensor. Lohr et al. used the DenseNet architecture for end-to-end eye movement biometrics (EMBs) as a new method for user authentication in VR and AR devices. An EER of 3.66% was achieved using 5 s of registration and authentication. However, because this study relies on high-quality eye movement data, eye-tracking sensors in VR/AR devices may have lower signal quality than the data used in this study.

2.3. User Authentication Using Periocular Authentication

Oishi et al. [21] proposed a method to improve mobile device user authentication by combining iris and periocular authentication using a machine learning algorithm called AdaBoost. To overcome the limitations of low-quality cameras commonly installed in mobile devices, this paper proposed using peripheral authentication in combination with iris scanning. This method compensated for the loss of iris authentication accuracy caused by low-resolution cameras. However, the method of this study relies heavily on the quality of the camera built into the mobile device. Additionally, an excellent image sensor and lens are required to obtain high-quality iris images. Zhao et al. [22] proposed a new framework for efficient and accurate matching of automatically acquired periocular images in a less restricted environment. They explained that by using the SCNN framework for periocular recognition, it can be useful in situations where accurate iris recognition is difficult. As a result of experiments on four databases, higher accuracy and EER were achieved compared to existing state-of-the-art methods. However, in this study, real-world data can often be highly variable and noisy, so a dataset that does not require preprocessing for these cases was used.

3. Data Acquisition

3.1. Ethics Statement

This study was exempt from review by the Sangmyung University Institutional Review Board as it did not involve any direct interaction with human subjects or any procedures that posed more than minimal risk to participants (IRB Exemption Number: EX-2023-006). Prior to the experiment, participants were provided with a detailed explanation of the experimental procedure and precautions, and informed consent was obtained. All personal data (e.g., name, periocular images) were collected anonymously. Additionally, the consent form stated that participants could withdraw from the experiment at any time if they felt dizziness or discomfort.

3.2. Participants

Through the distribution of flyers within the university and utilizing social networks, we recruited a total of 101 participants. The eligibility criteria for participation were healthy adults aged 18 and above. Participants with suboptimal vision were allowed to wear transparent contact lenses or glasses. However, for individuals wearing glasses, the experiments were conducted twice (once with glasses on and once without glasses). During the experiments, one participant (P057) requested to discontinue the experiments, and consequently, the data from this participant were excluded from the final dataset.

3.3. Apparatus

For the playback of virtual reality content, the HTC Vive Pro was used, featuring dual active-matrix organic light emitting diode (AMOLED) 3.5″ displays with a resolution of 1440 × 1600, a refresh rate of 90 Hz, and a field of view of 110°. Inside the HMD, an infrared camera from Pupil Labs, the HTC Vive Binocular Add-on, was installed to capture eye movements. This camera boasts a resolution of 1920 × 1080, 30 fps, a field of view over 100°, and a camera latency of 8.5 ms. Each camera is equipped with five infrared light emitting diodes (LEDs) to capture images of the eyes in dark environments. The cameras are connected to a laptop via USB cable, which has a 10-core 2.80 GHz i7-1165G7 CPU, 16 GB memory, and Intel(R) Iris(R) Xe Graphics. To display augmented reality content, the HMD is connected to a more powerful laptop with a 14-core 2.7 GHz i7-12700H CPU, 32 GB memory, and an NVIDIA RTX 3080 Laptop GPU with 16 GB VRAM. Although the HTC Vive Pro has a refresh rate of 90 Hz, the images are captured at 30 fps, resulting in a mismatch of frame rates that introduces noise into the video.

3.4. Procedures

To consider variations in pupil size and position related to emotions, as well as changes in pupil size influenced by image brightness, seven distinct images were carefully chosen for the experimental set. These images encompassed a range of valence and arousal values and varying levels of brightness. The video employed in a previous study [23] was a 360° video designed for the VR environment, providing valence and arousal values associated with each video. Four videos, representing positive arousal, positive non-arousal, negative arousal, and negative non-arousal, were selected from the aforementioned study. However, it was judged that the videos that induced positive and negative arousal were not appropriate, so through an internal meeting, external positive and negative arousal videos were added. Additionally, to induce neutral emotions, a method involving the display of everyday objects on the screen was adopted—a recognized approach for inducing neutral emotions [24]. Consequently, a total of seven videos were chosen, and Table 2 outlines the valence and arousal values, video length, and triggering emotion of each video. For externally sourced videos, the anticipated valence and arousal values were provided. The video set comprised two positive arousal videos, two negative arousal videos, one negative non-arousal video, one positive non-arousal video, and one neutral video. Figure 1 illustrates a example from the experiment's video. The images were thoughtfully

curated to encompass indoor and outdoor scenes, dark and bright environments, aiming to comprehensively consider changes in pupil size due to brightness.



Figure 1. Example from the experiment's video.

Each video has a duration ranging from 70 (s) to 90 (s), followed by a 3 min break after its conclusion to alleviate any potential dizziness. During this break, participants underwent a self-assessment using the self-assessment manikin (SAM) to gauge emotional arousal [25]. SAM is a widely employed method for investigating an experimenter's emotional response to various stimuli. The SAM includes a positive–negative scale representing emotions and an arousal–non-arousal scale indicating the level of arousal, as depicted in Figure 2, with each scale having nine levels. In addition, because this dataset assumes an actual usage environment, the user's free movement and gaze movement are not controlled, and since the HMD is worn again for each video, there is variation in eye position for each video even though it is the same person.

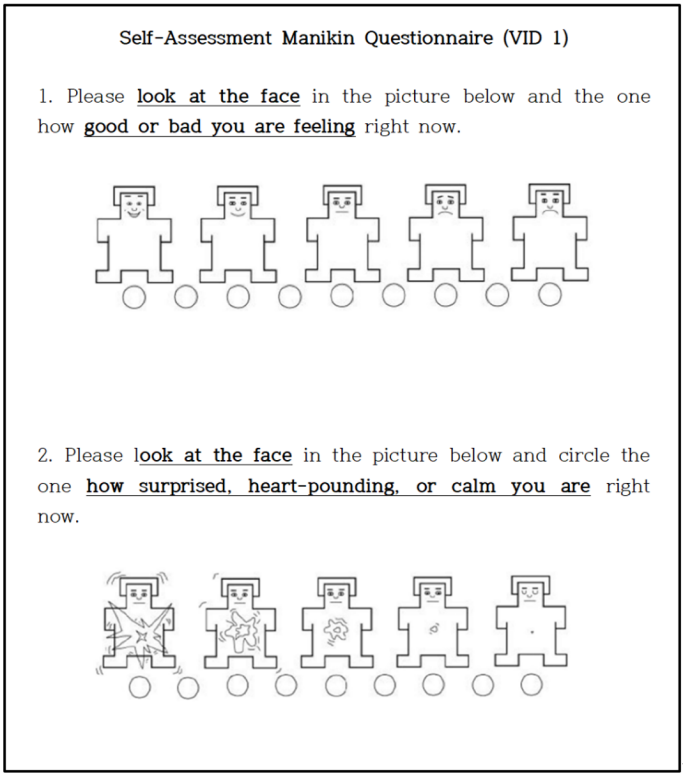


Figure 2. Self-assessment manikin questionnaire.

Table 2. Information about the video.

	VID 1	VID 2	VID 3	VID 4	VID 5	VID 6	VID 7
Valence	3.5 (expect)	7.47	5 (expect)	6.17	3.2	2.38	7 (expect)
Arousal	7 (expect)	5.35	3 (expect)	7.17	5.6	4.25	7.5 (expect)
Time (s)	90	70	90	90	90	90	90
Emotion	Negative arousal	Positive non-arousal	Neutral	Positive arousal	Negative arousal	Negative non-arousal	Positive arousal

3.5. Data Records

In this research, video acquisition was successfully completed for 100 out of the 101 participants, with one individual opting out of the experiment. Consequently, data for approximately 5,199,175 frames were obtained. The participants viewed seven videos designed to evoke emotions, and corresponding survey results were collected. The composition of the dataset is shown in Figure 3. Accordingly, the video data of each subject who participated in the experiment was structured to match the survey results for the video. During the experiment, information on whether lenses were worn and whether glasses were worn was also collected as metadata. In particular, whether the experiment was conducted first with glasses on or without them was expected to have a significant impact on emotion recognition, and the relevant information was recorded together. The resulting dataset was constructed by considering various aspects of the experiment and the reliability.

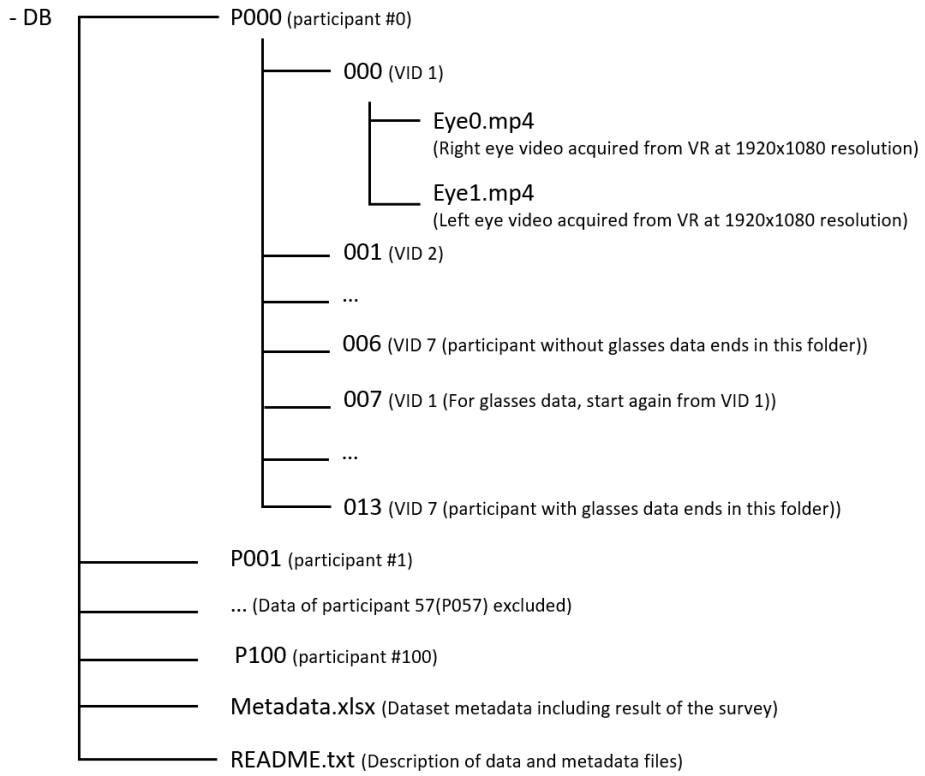


Figure 3. Composition of the dataset.

Figure 4 visually presents the data obtained through the experiment, and there is a change in the eye position of the same subject in each image. Furthermore, it clearly illustrates differences in periocular characteristics among subjects, including variations in eye shape, the presence or absence of double eyelids, and the shape of eyelashes. The observed distinctions in periocular features suggest the potential for subject identification based on these characteristics. It is worth noting that when subjects wear glasses, accurately

locating the eye area may pose challenges due to the reflection of infrared light by the lenses. Addressing these challenges may require additional image processing technology or controlled lighting conditions.

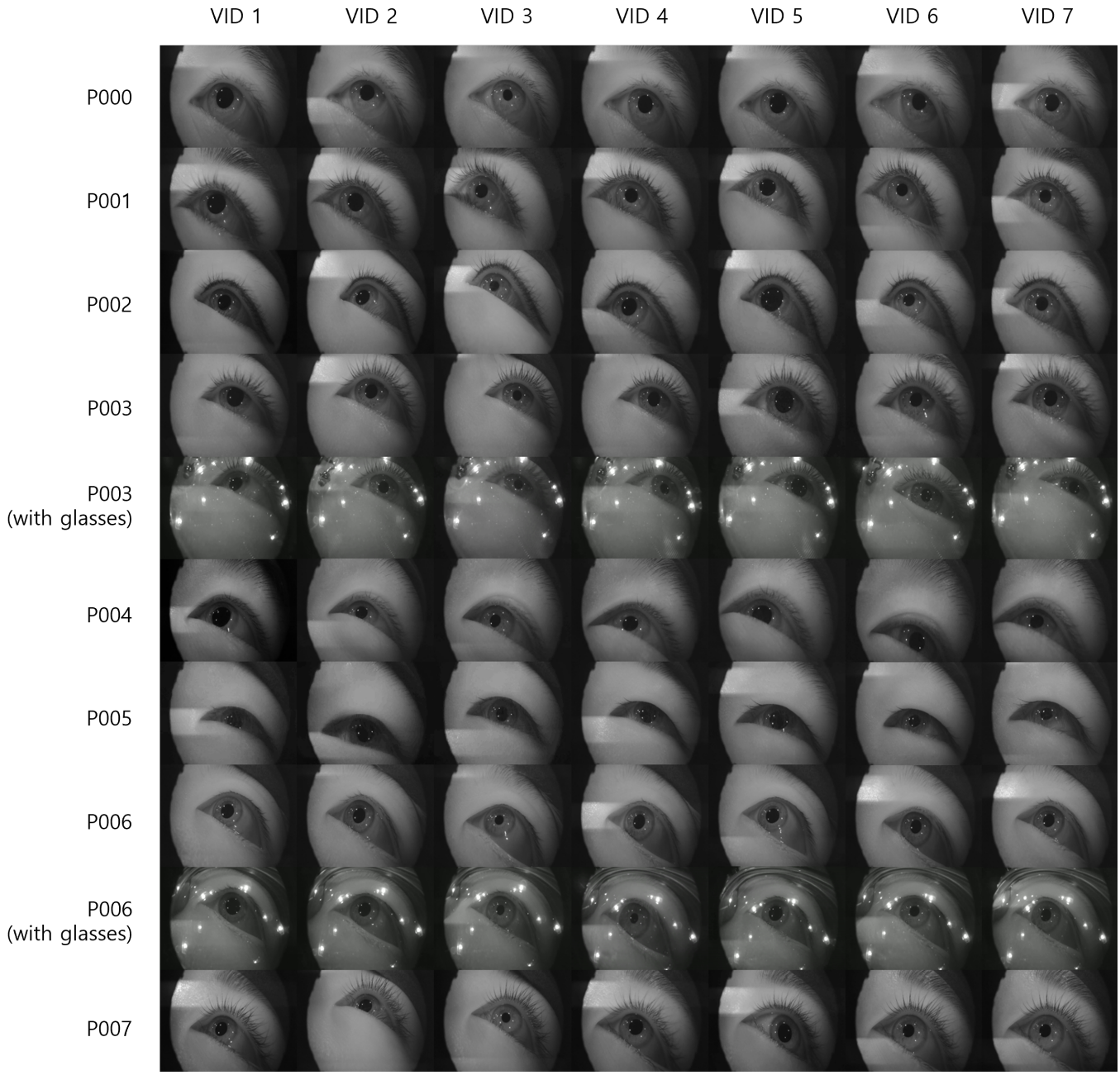


Figure 4. Data example (data from p000 to p007).

Figure 5 displays the image and histogram without glasses in (a) and with glasses in (b). In both cases, there is a substantial distribution of pixels with low brightness values (10–30) due to the VR-induced black borders around the eyes. Particularly in case (b), there is a significant increase in pixels with high brightness values (240 250) attributed to light reflecting from the glasses lens compared to (a). Moreover, the presence of glasses creates a shadow, and sporadic light-colored pixels exist due to dust or light reflection on the lenses, resulting in a distinctly different histogram. Additionally, in (a), distinct user features such as the iris and eyelashes are clearly visible in the center of the image, but in (b), these features are less evident as the pupils are notably positioned above the center due to the

presence of glasses. However, in both dataset types, the diameter of the iris in the entire periocular frame is not sufficient, so these datasets may not be suitable for iris recognition.

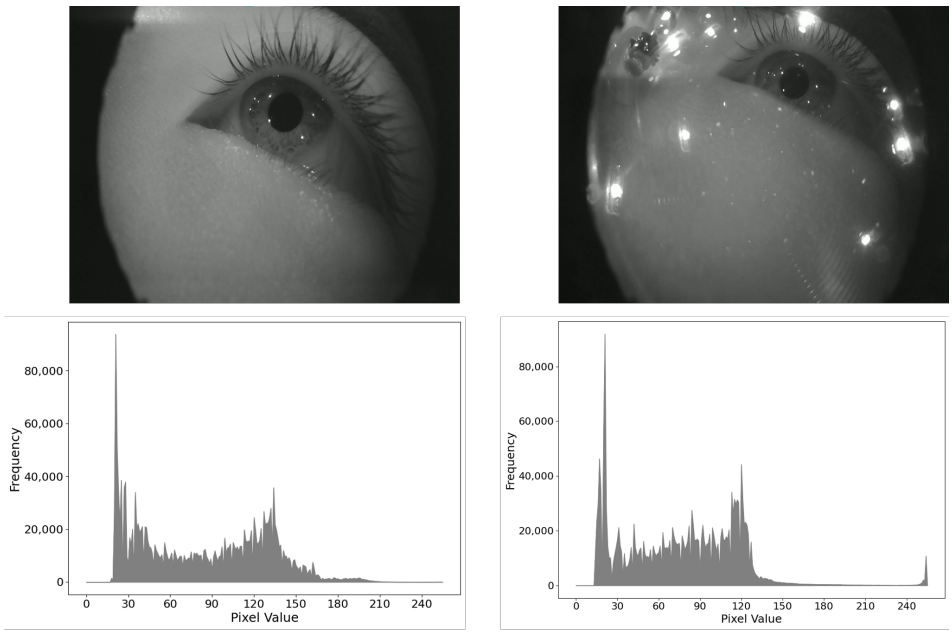


Figure 5. Example image (P003) and histogram. (**a**) Image w/ glasses; (**b**) image w/o glasses.

Figure 6 provides a visual representation of the evaluation results for each video, while Figure 7 illustrates the valence and arousal values of each image in a box-and-whisker plot. In Figure 6, you can see that the distribution of emotional levels for each video allows for relative distinction. However, it is evident from Table 3 that the valence and arousal values are denser than anticipated. Particularly, even in videos designed to induce arousal, the arousal values appear to be distributed lower, and overall valence values are distributed higher. Observing Figure 7, it is evident that valence demonstrates an overall distribution of high values, whereas arousal values exhibit an overall distribution of low values. This observation can be attributed to two potential factors. Firstly, during the selection of experimental videos, it is conceivable that the videos were chosen to elicit an overall low arousal value, possibly to accommodate a broader audience. Secondly, the immersive nature of wearing a VR headset and experiencing a 360-degree video may have induced positive emotions in the subjects.

Table 3. Average of valence and arousal as a result of the survey.

VID 1		**VID 2**		**VID 3**		**VID 4**	
Valence	Arousal	Valence	Arousal	Valence	Arousal	Valence	Arousal
4.12	5.33	7.6	2.54	5.61	1.61	6.12	4.79
VID 5		**VID 6**		**VID 7**			
Valence	Arousal	Valence	Arousal	Valence	Arousal		
4.33	4.79	3.88	2.79	7.59	4.15		

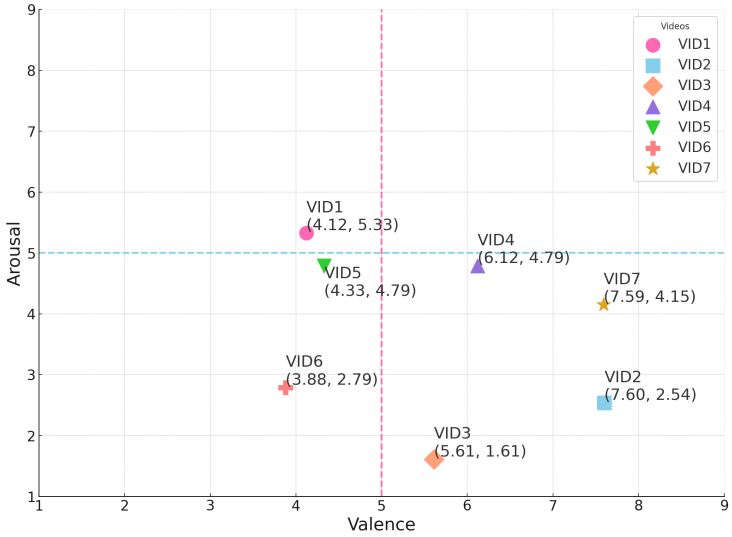


Figure 6. Survey results of valence and arousal distribution.

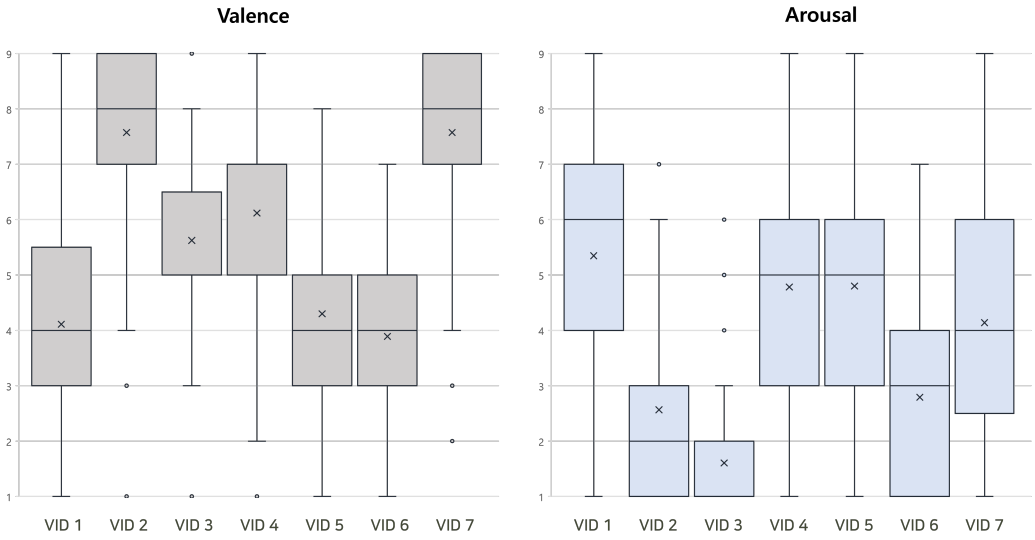


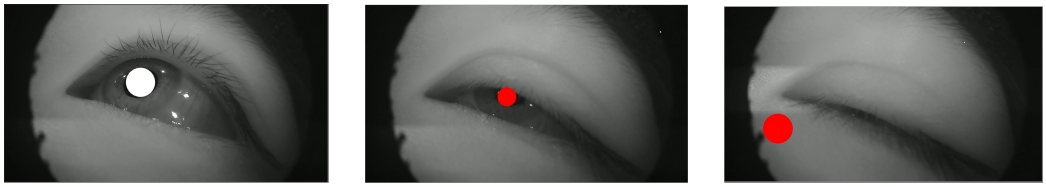
Figure 7. Box-and-whisker plot of valence and arousal as a result of the survey.

4. Method

4.1. Data Cleansing

Considering that the shape of both eyes is different for each person, only the right eye image was utilized in this study. Additionally, images featuring subjects wearing glasses were excluded from analysis due to the challenges posed by infrared light reflection on the lenses, making accurate eye area localization difficult. Moreover, images capturing closed or blinking eyes acquired during the experiment were omitted from consideration as they were deemed unsuitable for biometric recognition. The acquisition of eye images using infrared lights and cameras within the VR device introduce potential damage to iris images due to reflected infrared light in the iris area. To mitigate this, a process was implemented to remove reflected light present in the iris area. For the iris recognition task, an additional step involved converting the iris area into a square image using polar coordinate transformation. Given the dynamic nature of each user's eye position when wearing VR, a deep learning approach for accurate pupil extraction was deemed more effective than traditional algorithmic methods. Thus, a deep learning model based on Inception [26] was employed. Trained to detect and locate the pupil in real time within noisy images, this model consists of several inception blocks and reduction blocks, utilizing convolution filters and pooling layers of various sizes within each block. The model used

was pretrained, and to exclude instances of closed or slightly closed eyes, eight frames of video were removed based on the absence of detected pupils during the extraction process. Figure 8 visually illustrates the functioning of the model, with white and red circles indicating the predicted pupil model generated by the pupil area detection model. Frames classified as having open eyes are represented by white circles, while frames identified as closed eyes are visualized as red circles.



(a) (b) (c)

Figure 8. Examples of pupil extraction detection. A red dot indicates that the pupil was not detected in that frame, while a white dot indicates that the pupil was detected. Only frames where the pupil was detected were used for training the model. (**a**): When the eyes are open, (**b**) when eyes are half-open (in this case, it is classified as a frame with eyes closed), (**c**): when eyes are closed.

4.2. Dataset

For training the biometric recognition model, images were extracted and utilized from our dataset. Participants in this database viewed a total of seven videos, and from each, 15 images were randomly extracted for the training process. A total of 105 images were used per subject, and a total of 10,500 learning data for 100 subjects were constructed and model learning was performed. Considering the variability of the VR wearing environment, random parallel movement and brightness augmentation were applied to the images when learning the model. Additionally, considering that impostor pairs are more diverse than genuine pairs in an actual biometric authentication environment, the ratio of impostor pairs was increased when forming genuine and impostor pairs in 10,500 pieces of training data to enable robust comparison. The genuine and impostor matching tests were conducted using a dedicated test dataset, excluding the data utilized for training. Considering the relatively small number of subjects in the database, five-fold cross-validation performance was measured. The overall average performance was then calculated to evaluate the system.

4.3. Periocular Recognition Model

In this study, a Siamese-network-based deep learning model was utilized to compare and analyze the performance of biometric recognition models based on periocular data obtained from a VR environment. The Siamese network is a deep learning model structure that shares weights between CNN models. In the case of genuine image pairs, the feature values of the two images become closer, while in the case of impostor pairs, the feature values become further apart. The overall model structure is shown in Figure 9. In this study, the performance comparison was conducted by modifying the feature extraction network model based on the Siamese network. The three deep learning models used as feature extraction networks were MobileNetV3Large [27], EfficientNetB0 [28], and the Siamese-network-based deep learning model proposed by Hwang et al. [29]. Each model has its unique characteristics, covering a wide spectrum from lightweight models to high-performance models.

MobileNetV3 [27] is a lightweight deep learning model designed for efficient computational performance in mobile and embedded environments. The MobileNet series utilizes depthwise separable convolution, which reduces the number of parameters while optimizing the computation speed. MobileNetV3 further advances this by introducing the squeeze-and-excitation (SE) module and the hard swish activation function, striking a balance between performance and efficiency. It is evaluated as a model that consumes fewer computational resources, performs fast computations, yet still delivers respectable

recognition performance. In this study, MobileNetV3Large was selected, considering the possibility of real-time processing on VR devices.

EfficientNet [28] is a model designed with a compound scaling strategy, which expands the network's depth, width, and resolution in a balanced way. EfficientNetB0 is the lightest version in this series, balancing performance and efficiency, making it a model that can be applied to various use cases. A key feature of EfficientNet is that it does not simply improve performance by increasing the model's size but achieves optimization in both performance and efficiency through compound scaling. EfficientNetB0's strength lies in its ability to deliver high performance even with limited data, making it suitable for the VR dataset, which was collected from a relatively small number of subjects.

The study by [29] argued that useful features for periocular biometrics might exist even at relatively lower layers in CNN-based models. As a result, the model was designed to utilize features extracted from intermediate layers of the deep learning model. Based on Deep-ResNet18, feature maps extracted from each stage are transformed into vectors using Global Average Pooling. These transformed vectors are then connected to a fully connected layer, mapping them to vectors of the same size. Consequently, one vector is generated for each convolutional layer stage, and the final feature vector is created by concatenating these vectors. This feature vector is used to perform periocular biometrics by comparing the two subjects.

The input images used for the model had a resolution of (240, 320). All models were trained using the TensorFlow and Keras libraries, and the Adam optimizer, which generally shows good performance, was used as the optimizer for training. The binary cross-entropy was employed as the loss function. To save the model with the best performance, the model parameters at the point where the validation loss was the lowest were saved. The batch size for training was set to 16, and the model was trained for a total of 10 epochs. Additionally, to prevent overfitting during training, the early stopping technique was applied, halting training if the validation loss did not improve for five consecutive epochs.

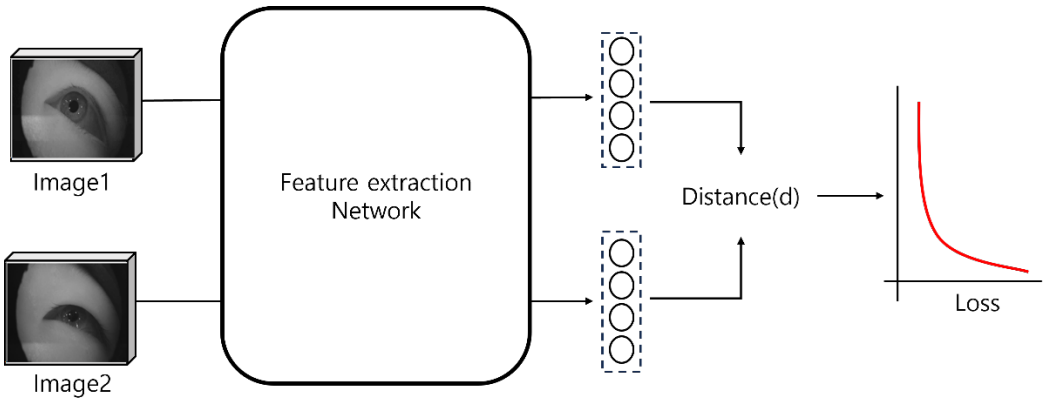


Figure 9. Siamese network structure for model training.

5. Results

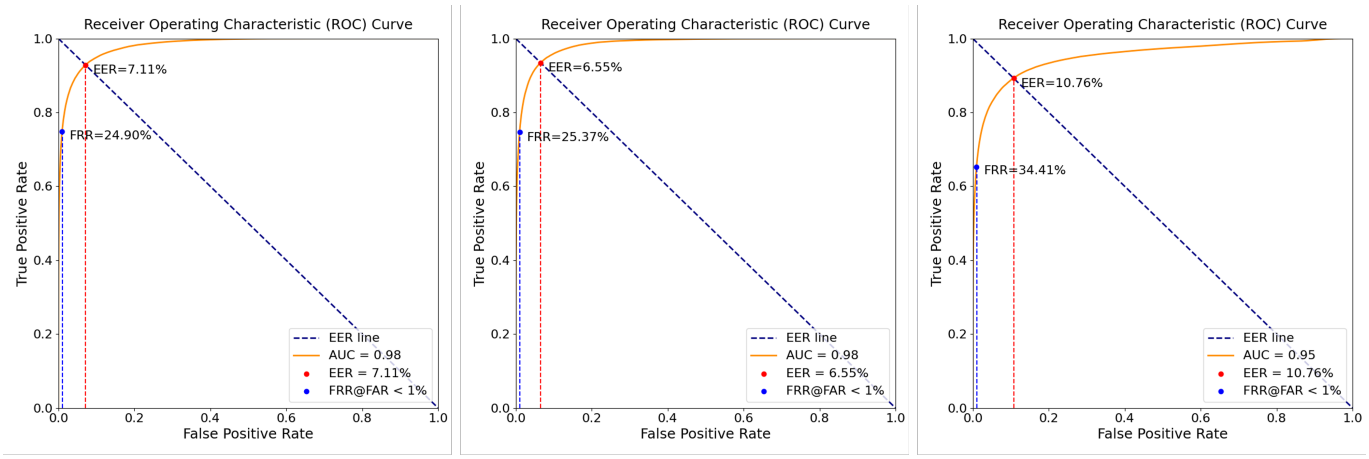
In this study, we measured the false acceptance rate (FAR) and false rejection rate (FRR) based on similarity thresholds to compare the performance of the models and conducted receiver operating characteristic (ROC) curve and equal error rate (EER) analyses. A lower EER indicates a more robust biometric recognition system, and in most biometric systems, the threshold is typically set based on the value derived from the EER. In biometric recognition, minimizing the FAR, which is associated with the incorrect identification of others, is critical. Consequently, the FRR performance was evaluated under the condition that FAR remained below 1%. Table 4 summarizes the performance of the proposed biometric recognition models. As seen in the table, the performance of each model varies in terms of EER and FRR. MobileNetV3Large and EfficientNetB0 exhibited similar performances, with EERs of 7.11% and 6.55%, respectively. In contrast, the model proposed

by Hwang et al. [29] showed relatively lower performance with an EER of 10.76%. When comparing FRR at a FAR of less than 1%, MobileNetV3Large and EfficientNetB0 achieved similar performance, with FRRs of 24.90% and 25.37%, respectively, whereas the model proposed by Hwang et al. [29] displayed a lower performance with an FRR of 34.41%.

Table 4. Biometric recognition performance of different models.

Model	EER [%]	FRR (FAR < 1%) [%]
MobileNetV3Large	7.10	24.90
EfficientNetB0	6.55	25.36
Hwang et al. [29]	10.76	34.41

Figure 10 visualizes the ROC curves for each model. Through the analysis of each curve, it was found that the MobileNetV3Large model demonstrated an AUC of 0.98, which is very close to 1, indicating excellent performance. Similarly, the EfficientNetB0 model also showed an AUC of 0.98, exhibiting identical performance to MobileNetV3Large and demonstrating very high performance. Given that EfficientNetB0 has a lower EER than MobileNetV3Large, both models display similar overall performance, but EfficientNetB0 may slightly outperform MobileNetV3Large. On the other hand, the model proposed by Hwang et al. [29] achieved an AUC of 0.95, which, although slightly lower than the other two models, still maintains high performance.



(a) MobileNetV3Large **(b) EfficientNetB0** **(c) Hwang et al. 2020**

Figure 10. Comparison of ROC curves: (**a**) MobileNetV3Large; (**b**) EfficientNetB0; (**c**) Hwang et al. [29].

Figure 11 illustrates the genuine–impostor distribution of the models using the proposed dataset as input. The X-axis represents the distance between the feature vectors of two images, while the Y-axis represents the probability density at that distance. The closer the feature vector distance between two images, the higher the likelihood that the pair is genuine, and the further the distance, the higher the likelihood that the pair is an impostor. Typically, the distributions of genuine and impostor pairs each form two Gaussian distributions, and the smaller the overlapping region between the two distributions, the better the biometric recognition performance is considered to be. For the MobileNetV3Large model, genuine data are primarily located between Euclidean distances of 0.0 and 0.8. The distribution of the genuine data is very narrow and skewed to the left, indicating that genuine matches occur at very small distances. The impostor data are mainly positioned between Euclidean distances of 0.2 and 1.5 and follow a symmetric normal distribution. This suggests that the MobileNetV3Large model can effectively distinguish between genuine and

impostor data, as the overlap between the two distributions is relatively small, indicating excellent performance. In contrast, the EfficientNetB0 model shows a broader distribution for genuine data compared to MobileNetV3Large, with the distribution spreading further to the right. The genuine data is primarily located between Euclidean distances of 0.0 and 1.0. And the impostor data is mainly positioned between Euclidean distances of 0.2 and 1.8 and follows a symmetric normal distribution. The impostor distribution also spreads more to the right, forming a symmetrical distribution. This model's broader distributions for both genuine and impostor data imply that the model includes more uncertainty, and the boundary between the two classes is relatively less distinct. Therefore, despite having the lowest EER, the EfficientNetB0 model suggests that its criteria for distinguishing between genuine and impostor pairs are not as clear as those of the MobileNetV3Large model. For the model proposed by Hwang et al. [29], the genuine distribution is more spread out compared to the previous two models, with a slightly asymmetrical normal distribution skewed to the left. The impostor distribution is similar to that of the EfficientNetB0 model but is skewed more to the left. The genuine data are spread between Euclidean distances of 0.2 and 1.5, and there is more overlap between the distributions of genuine and impostor data than in the previous two models. This indicates that the difference between genuine and impostor data is smaller, which contributes to the relatively lower performance of this model.

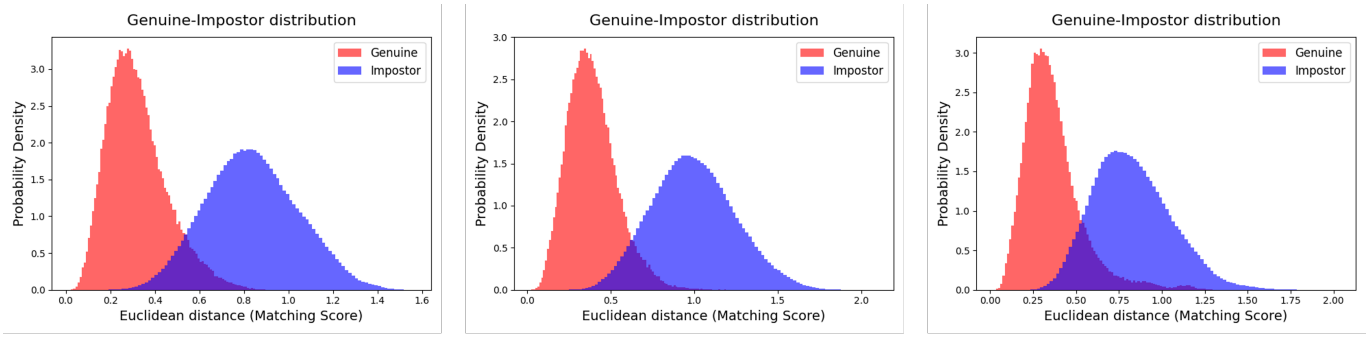


Figure 11. Comparison of genuine–impostor distribution: (**a**) MobileNetV3Large; (**b**) EfficientNetB0; (**c**) Hwang et al. [29].

6. Discussion

This study attempted to overcome the limitations of existing datasets by building a VR dataset recorded while subjects experienced actual VR content. Most existing public datasets were acquired from Westerners and were not suitable for VR research because they did not reflect the uniqueness of the VR environment by using data acquired from external cameras. Additionally, datasets captured in VR have a small number of subjects or low resolution. In contrast, this study collected high-resolution images using a camera attached to a VR device targeting Koreans and constructed a richer dataset through emotion-inducing videos. To induce emotions, the subjects watched seven videos, and after watching, the subjects checked the level of positive/negative and arousal/non-arousal through a self-assessment manikin questionnaire. The results of the questionnaire showed that the valence value was generally high, and the arousal value was generally distributed similarly across all videos. This may be because, during the selection process for the experimental videos, videos that generally induce low arousal values were included, which various subjects could watch. It is also possible that this resulted from cultural differences between Western and Eastern societies. It may also be because the experience of wearing VR and watching 360-degree videos itself induced positive emotions in the subjects. In the case of emotion recognition, it is feasible to predict the numerical values corresponding to emotions in the videos used in our constructed dataset through regression, using metrics

such as mean absolute error, or to utilize each emotion as a category and use accuracy for prediction.

This study also presents a baseline for utilizing the dataset. Periocular biometric identification was performed using the constructed dataset. To improve the quality of the data, frames with eyes closed or blinking were removed, and frames with rapid changes in pupil size were filtered. This process improves the reliability and accuracy of biometrics, allowing the model to effectively learn the key features needed to distinguish between real and counterfeit. And the feature extraction network of the Siamese network was replaced with various models for comparative analysis. The results of this study highlight the comparative performance of three feature extraction networks—MobileNetV3Large, EfficientNetB0, and the model proposed by Hwang et al. (2020)—in terms of EER, FAR, and FRR. The models were evaluated using ROC curves and genuine–impostor distribution analysis to gauge their accuracy and ability to distinguish between genuine and impostor data. The analysis of the genuine–impostor distribution further illuminates the strengths and weaknesses of each model. The MobileNetV3Large model presents a compact and well-separated distribution of genuine and impostor data, with genuine data clustering tightly between Euclidean distances of 0.0 and 0.8. This narrow distribution, coupled with minimal overlap between the two classes, indicates that MobileNetV3Large can effectively distinguish between genuine and impostor matches, making it highly reliable for biometric recognition. In contrast, EfficientNetB0, despite its lower EER, shows a broader distribution of both genuine and impostor data. This broader spread suggests more uncertainty in the model's classification boundaries, making it less clear-cut in distinguishing between genuine and impostor data. The overlap between the distributions is larger than that of MobileNetV3Large, indicating that EfficientNetB0, while effective, may not be as confident in its predictions.

Also, balancing FAR and FRR is a critical challenge in biometric recognition systems. In this study, we evaluated FRR under the condition that FAR remained below 1%, as well as by analyzing the EER. However, there is a need for more detailed analysis of FRR performance at various FAR thresholds. Specifically, exploring methods to achieve an optimal FAR/FRR balance tailored to the needs of specific applications could be a key focus for future research. Also, no emotion rating data were used in the baseline, and images of glasses wearers were excluded from the analysis due to infrared reflection. Therefore, future research will expand the scope of application of the model by including various states such as glasses wearers, eye closure, and blinking and will contribute to expanding the use of periocular in more diverse fields through emotion classification, etc.

7. Conclusions

This study establishes a periocular dataset acquired in real VR usage environments, providing a crucial foundation for research in biometric recognition and emotion assessment. Existing datasets fail to reflect the unique characteristics of VR environments and often have limitations such as low resolution or a small number of subjects. In contrast, the dataset from this study includes high-resolution images captured using a camera attached to a VR device, specifically focusing on Korean subjects, making it more suitable for practical applications. The significance of this dataset is particularly highlighted in its potential to enable non-invasive continuous authentication and emotion assessment in VR environments. This opens possibilities for applications in fields where immersion is critical, such as education, therapy, and entertainment. The data collected alongside emotion-inducing videos can serve as a valuable resource for emotion recognition research, contributing to the development of more personalized and sophisticated VR environments. In conclusion, the dataset created in this study will serve as a key cornerstone for advancing technologies in user authentication and emotion recognition within VR environments. Based on this, the quality of user experiences can be enhanced, and safer, more reliable VR environments can be established. The expansion of such research lays the foundation for

VR technology to provide tangible value across various industries and make significant contributions to the future development of immersive technologies.

Author Contributions: Conceptulization—E.C.L.; Methodology—C.S.; Software—Y.P.; Validation—J.B.; Formal Analysis—C.S.; Investigation—J.-h.R., Y.K. and S.K.; Resources—C.S., Y.P., J.B. and H.L.; Data Curation—C.S.; Writing—Original Draft—C.S.; Writing—Review and Editing—E.C.L.; Visualization—C.S., Y.P. and J.B.; Supervision—E.C.L.; Project Administration—E.C.L.; Funding Acquisition—J.-h.R., Y.K. and S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00215700, Trustworthy Metaverse: blockchain-enabled convergence research).

Institutional Review Board Statement: Based on Article 13-1-1 and 13-1-2 of the Enforcement Regulations of the Act on Bioethics and Safety of the Republic of Korea, ethical review and approval were waived (IRB-SMU-C-2023-1-007) for this study by Sangmyung University's Institutional Review Board because this study uses only simple contact-measuring equipment or observation equipment that does not involve physical changes and does not include invasive procedures such as drug administration or blood sampling.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the subjects to publish this paper.

Data Availability Statement: All codes were released and shared at (https://github.com/schaelin/AffectiVR (accessed on 24 September 2024)). The code was written using Python 3.8.18, and how to use it is explained at (https://github.com/schaelin/AffectiVR (accessed on 24 September 2024)).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Funk, M.; Marky, K.; Mizutani, I.; Kritzler, M.; Mayer, S.; Michahelles, F. Lookunlock: Using spatial-targets for user-authentication on hmds. In Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019; pp. 1–6.
2. Rose, T.; Nam, C.S.; Chen, K.B. Immersion of virtual reality for rehabilitation-Review. *Appl. Ergon.* **2018**, *69*, 153–161. [CrossRef] [PubMed]
3. Kim, S.; Kim, S.; Jin, S. Trends in Implicit Continuous Authentication Technology. *Electron. Telecommun. Trends* **2018**, *33*, 57–67.
4. Kumari, P.; Seeja, K. Periocular biometrics: A survey. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 1086–1097. [CrossRef]
5. Alonso-Fernandez, F.; Bigun, J. A survey on periocular biometrics research. *Pattern Recognit. Lett.* **2016**, *82*, 92–105. [CrossRef]
6. Joo, J.H.; Han, S.H.; Park, I.; Chung, T.S. Immersive Emotion Analysis in VR Environments: A Sensor-Based Approach to Prevent Distortion. *Electronics* **2024**, *13*, 1494. [CrossRef]
7. Petersen, G.B.; Petkakis, G.; Makransky, G. A study of how immersion and interactivity drive VR learning. *Comput. Educ.* **2022**, *179*, 104429. [CrossRef]
8. Li, S.; Yi, D.; Lei, Z.; Liao, S. The casia nir-vis 2.0 face database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 348–353.
9. Kumar, A.; Passi, A. Comparison and combination of iris matchers for reliable personal authentication. *Pattern Recognit.* **2010**, *43*, 1016–1026. [CrossRef]
10. Bowyer, K.W.; Flynn, P.J. The ND-IRIS-0405 iris image dataset. *arXiv* **2016**, arXiv:1606.04853.
11. Proença, H.; Alexandre, L.A. UBIRIS: A noisy iris image database. In *Proceedings of the Image Analysis and Processing–ICIAP 2005: 13th International Conference, Cagliari, Italy, 6–8 September 2005*; Proceedings 13; Springer: Cagliari, Italy, 2005; pp. 970–977.
12. Proença, H.; Filipe, S.; Santos, R.; Oliveira, J.; Alexandre, L.A. The UBIRIS. v2: A database of visible wavelength iris images captured on-the-move and at-a-distance. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 1529–1535. [CrossRef] [PubMed]
13. Fusek, R. Pupil localization using geodesic distance. In *Proceedings of the Advances in Visual Computing: 13th International Symposium, ISVC 2018, Las Vegas, NV, USA, 19–21 November 2018*; Proceedings 13; Springer: Las Vegas, NV, USA, 2018; pp. 433–444.
14. Garbin, S.J.; Shen, Y.; Schuetz, I.; Cavin, R.; Hughes, G.; Talathi, S.S. Openeds: Open eye dataset. *arXiv* **2019**, arXiv:1905.03702.
15. Kagawade, V.C.; Angadi, S.A. VISA: A multimodal database of face and iris traits. *Multimed. Tools Appl.* **2021**, *80*, 21615–21650. [CrossRef]
16. Kim, J.; Stengel, M.; Majercik, A.; De Mello, S.; Dunn, D.; Laine, S.; McGuire, M.; Luebke, D. Nvgaze: An anatomically-informed dataset for low-latency, near-eye gaze estimation. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Scotland, UK, 4–9 May 2019; pp. 1–12.
17. Palmero, C.; Sharma, A.; Behrendt, K.; Krishnakumar, K.; Komogortsev, O.V.; Talathi, S.S. Openeds2020: Open eyes dataset. *arXiv* **2020**, arXiv:2005.03876.

18. Liebers, J.; Schneegass, S. Gaze-based authentication in virtual reality. In Proceedings of the ACM Symposium on Eye Tracking Research and Applications, Stuttgart, Germany, 2–5 June 2020; pp. 1–2.
19. Luo, S.; Nguyen, A.; Song, C.; Lin, F.; Xu, W.; Yan, Z. OcuLock: Exploring human visual system for authentication in virtual reality head-mounted display. In Proceedings of the 2020 Network and Distributed System Security Symposium (NDSS), San Diego, CA, USA, 23–26 February 2020.
20. Lohr, D.; Komogortsev, O.V. Eye know you too: Toward viable end-to-end eye movement biometrics for user authentication. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 3151–3164. [CrossRef]
21. Oishi, S.; Ichino, M.; Yoshiura, H. Fusion of iris and periocular user authentication by adaboost for mobile devices. In Proceedings of the 2015 IEEE International Conference on Consumer Electronics (ICCE), Las Vegas, NV, USA, 9–12 January 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 428–429.
22. Zhao, Z.; Kumar, A. Accurate periocular recognition under less constrained environment using semantics-assisted convolutional neural network. *IEEE Trans. Inf. Forensics Secur.* **2016**, *12*, 1017–1030. [CrossRef]
23. Li, B.J.; Bailenson, J.N.; Pines, A.; Greenleaf, W.J.; Williams, L.M. A public database of immersive VR videos with corresponding ratings of arousal, valence, and correlations between head movements and self report measures. *Front. Psychol.* **2017**, *8*, 2116. [CrossRef] [PubMed]
24. Trilla, I.; Weigand, A.; Dziobek, I. Affective states influence emotion perception: Evidence for emotional egocentricity. *Psychol. Res.* **2021**, *85*, 1005–1015. [CrossRef] [PubMed]
25. Lang, P.; Sidowski, J.; Johnson, J.; Williams, T. *Technology in Mental Health Care Delivery Systems*; Ablex Publishing Corporation: Norwood, NJ, USA, 1980.
26. Eivazi, S.; Santini, T.; Keshavarzi, A.; Kübler, T.; Mazzei, A. Improving real-time CNN-based pupil detection through domain-specific data augmentation. In Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, Denver , CO, USA, 25–28 June 2019; pp. 1–6.
27. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
28. Tan, M. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
29. Hwang, H.; Lee, E.C. Near-infrared image-based periocular biometric method using convolutional neural network. *IEEE Access* **2020**, *8*, 158612–158621. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.