electronics

MDPI

Article

Improving the Performance of Automatic Lip-Reading Using Image Conversion Techniques

Ki-Seung Lee

Department of Electronic Engineering, Konkuk University, 1 Hwayang-dong, Gwangjin-gu, Seoul 143-701, Republic of Korea; kseung@konkuk.ac.kr; Tel.: +82-02-450-3489

Abstract: Variation in lighting conditions is a major cause of performance degradation in pattern recognition when using optical imaging. In this study, infrared (IR) and depth images were considered as possible robust alternatives against variations in illumination, particularly for improving the performance of automatic lip-reading. The variations due to lighting conditions were quantitatively analyzed for optical, IR, and depth images. Then, deep neural network (DNN)-based lip-reading rules were built for each image modality. Speech recognition techniques based on IR or depth imaging required an additional light source that emitted light in the IR range, along with a special camera. To mitigate this problem, we propose a method that does not use an IR/depth image directly, but instead estimates images based on the optical RGB image. To this end, a modified U-net was adopted to estimate the IR/depth image from an optical RGB image. The results show that the IR and depth images were rarely affected by the lighting conditions. The recognition rates for the optical, IR, and depth images were 48.29%, 95.76%, and 92.34%, respectively, under various lighting conditions. Using the estimated IR and depth images, the recognition rates were 89.35% and 80.42%, respectively. This was significantly higher than for the optical RGB images.

Keywords: silent speech interface; speech pattern recognition; deep neural networks; infrared image; depth image; image conversion

check for updates

Citation: Lee, K.-S. Improving the Performance of Automatic Lip-Reading Using Image Conversion Techniques. *Electronics* **2024**, *13*, 1032. https://doi.org/10.3390/electronics13061032

Academic Editor: Włodzimierz Kasprzak

Received: 29 December 2023
Revised: 23 February 2024
Accepted: 24 February 2024
Published: 9 March 2024

Copyright: © 2024 by the author. 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

A voice is a very useful means of communication, but it cannot be used under conditions where the ambient noise level is very high or where security concerns make it necessary to conceal voice information. Silent speech interface (SSI) techniques [1] were proposed to cope with these situations. In an SSI, a voice is synthesized using non-speech modalities that are then moderately correlated with the voice. Images acquired of a speaker's mouth region are widely used in an SSI, and the underlying principle is that the shape of the mouth is determined by the speech being uttered. Speech recognition using an image modality (so-called "automatic lip-reading" [2,3]) was initially applied to simple digit- [4–7]/alphabet-recognition [8,9] problems, and it then was applied to isolated word recognition [10–13]. A sub-word (e.g., phoneme)-based recognition scheme by means of live videos was proposed to implement a sentence-level recognizer [14–16]. A recognition rate of 88.5% was achieved in isolated word recognition [13]. For sentence-level recognition, 64.6% was achieved when three-dimensional convolutional neural networks were adopted [14]. Such results demonstrate the possibility of identifying the contents of speech being uttered using only images of the corresponding speaker, without any audio signal.

One of the fundamental drawbacks associated with image-based pattern recognition is that the recognition accuracy is significantly affected by discrepancies in the reference pattern due to differences in acquisition conditions. Automatic lip-reading is often used indoors, and therefore, in artificial rather than natural light. As a result, there is a high probability of fluctuations in the image depending on the lighting environment. Variation in lighting conditions is a major cause of such discrepancies [17–19]. Intensity

normalization [20], contrast enhancement [21,22], histogram equalization [23,24], and image dehazing [25,26] can be applied to cope with this problem. By using such methods, however, performance is limited under nonuniform lighting or extremely bright or dark lighting. Furthermore, distortions caused by excessive or incorrect image processing can further reduce the recognition accuracy.

The usage of depth images or infrared (IR) images can be an alternative solution for the problems associated with uneven lighting conditions [27–31]. Since such images are acquired under invisible IR light rather than visible light, this option has the advantage of being able to obtain meaningful images, even in a dark environment that is not affected by ambient lighting. The depth of an image was used to recognize hand gestures under various backgrounds and lighting conditions [32] that were obtained using a time-of-flight (TOF) camera or stereo camera. In the case of using a depth camera, however, the issues associated with cost and low resolution compared with using a conventional RGB camera should be addressed. In the case of acquiring an IR image, it is possible to manufacture a light source using inexpensive IR LEDs. Moreover, a typical RGB camera with a silicon image sensor is capable of acquiring near-infrared (NIR) images with a relatively high quality and with a high level of the signal-to-noise ratio at a wavelength of around 800 nm [33]. Due to these advantages, IR images are being adopted in various applications, including face detection/recognition [30], action recognition [31], and human motion analysis [34]. A thermal image is a type of IR image that is acquired without the influence of ambient visible light. The characteristics of thermal images were used to recognize expressions in facial images via fusion with visible light images [35].

In this study, the usefulness of depth and IR images was verified from the perspective of automatic lip-reading, particularly when the lighting conditions were very different from reference sources. For three types of images (optical, IR, and depth images), variations due to lighting conditions were quantitatively measured using images captured from the speaker's mouth region. Then, speech recognition was carried out under various lighting conditions for 80 isolated words, wherein three-dimensional discrete Fourier transformation (3D DFT) and deep neural networks were employed for the construction of the classification rules. The experimental results verified that IR and depth images could reasonably maintain a recognition accuracy that approximated results from conventional optical images, even under various lighting conditions.

Despite the advantage of IR/depth imaging reducing the effects of lighting conditions, there are some challenges associated with its practical implementation. First, a separate light source that emits IR light is required to acquire images in the IR band. When using a TOF camera, a light source in the IR band is also required. Second, a separate camera or additional acquisition time is required to acquire IR images, which increases the acquisition time compared with using optical RGB images alone. Furthermore, obtaining depth images requires complex computational processes using the acquired optical and IR images, which often increases the complexity of automatic lip-reading devices. In this study, we focused on how to mitigate the challenges caused by IR/depth imaging acquisition while retaining the advantages that are not affected by lighting conditions.

Previous work investigated the use of RGB images to estimate other types of images, such as depth distribution [36–40] and semantic segmentation mapping [41–43]. This was based on the assumption that the boundary information in an RGB image corresponds in some way to the in depth distribution and that the intensity of the reflected light is inversely proportional to the distance between the camera and the object. However, it is not easy to explain the relationship between the RGB image and its depth distribution. Accordingly, the depth distribution is predicted mostly from the RGB image by using a nonlinear regression model, such as a convolutional neural network (CNN) [38–40]. In previous studies, the estimation of depth images was applied mainly to indoor and outdoor images [37–40] and showed very promising results. Based on these findings, for use in lip-reading, we adopted a method where depth images were estimated from RGB images instead of actually being acquired. To this end, two issues should be considered: (1) Compared with typical indoor

and outdoor images, the facial images for lip-reading have a shorter camera-to-subject distance and a finer depth resolution. (2) Unlike previous studies that were focused on estimating the precision of the depth image itself, the depth images should be estimated from the perspective of speech recognition accuracy.

An attempt was also made to use RGB images to predict IR images for use in visualizing vegetation mapping [44–47] and vein patterns [48–50]. In order to improve the accuracy of lip-reading under different lighting conditions, RGB-to-IR image conversion was also adopted in this study. The uniqueness of this proposed method lies in the fact that a regression analysis approach was employed to correct images distorted by lighting, and images free from lighting conditions were used as reference images. Unlike traditional contrast-correction techniques, where the visual quality of the image is a major concern, the goal of this work was to improve the accuracy of automatic lip recognition. Therefore, it was important to ensure that the processed image was not close to an optical image, but rather an infrared/depth image, which was less affected by lighting conditions.

The outline of this paper is as follows: After the introduction, Section 2 compares three types of images (optical, NIR, and depth) for the amount of variation due to changes in lighting conditions. Section 3 provides an overview of the proposed visual speech recognition method. Section 4 describes how NIR and depth images were estimated using optical RGB images. Section 5 provides the results from several tests, including the recognition accuracies of each modality in the cases of using both actual and predicted images. Section 6 concludes this work and suggests future directions.

2. Comparison of Optical/IR/Depth Images

Prior to the construction of the classification rules for automatic lip-reading, mouth images were captured using RGB, IR, and depth cameras under various lighting conditions, and then variations according to lighting conditions were observed for each image modality. A webcam-type RGB camera equipped with IR filters was employed to capture the optical and IR images. The optical and IR images were selectively acquired using a motorized IR filter ($\lambda_{cut} \approx 700$ nm) and two types of lighting sources. The illumination used to acquire the optical and IR images was configured using white LEDs and IR LEDs, respectively. Various lighting conditions were produced by changing the distance between the LED light and the subject's face, the direction of the LED light, and the driving current. A TOF camera was employed to obtain the depth images. A video was recorded using one male subject who was asked to pronounce the five Korean vowels /a:/, /i:/, /u:/, /e:/, and /o:/. For each vowel, a still image was captured at the midpoint of the vowel segment.

Examples of the mouth images (when voicing a phoneme /a:/) were captured under normal, dark, nonuniform, and bright lighting conditions, as shown in Figure 1. There was a large visual difference in the optical images according to the lighting conditions. These differences were compensated for using image-processing techniques, such as intensity normalization and histogram equalization. The processed images also appear in Figure 1. At very low-intensity levels, the compensated images seemed noisy. Such artifacts imposed adverse effects on the recognition accuracy. The nonuniformity of the images due to uneven lighting was not sufficiently corrected by the intensity compensation techniques, as shown in the third column of the figure.

The IR images shown in Figure 1 were taken using a typical RGB camera and a light source with a wavelength of 870 nm. The IR images were acquired under a light source with wavelengths outside the visible range, but showed good morphological agreement with the optical images. Moreover, since visible light was blocked by a filter during the capture period, the obtained images were not affected by illumination from either visible or ambient light. Compared with the optical/IR images, the image details were not observed and only the approximate shape of the mouth was seen in the depth images. Nevertheless, it was possible to capture the images without the affects of ambient visible light.

In this study, the variations according to the lighting conditions for each image modality were both visually examined and quantitatively measured. The distribution of the pixel

values (histogram) was used to measure the difference between two images acquired under different lighting conditions. A histogram was obtained from the lower region of the face in the captured images. The Bhattacharyya distance [51] was employed to measure the difference between images. The Bhattacharyya distance for two images I_n and I_m acquired under lighting conditions c_n and c_m, respectively, is given by

$$D_B(I_n, I_m|c_n, c_m) = -\log\left\{\sum_{y\in Y}\sqrt{p(y|I_n, c_n)p(y|I_m, c_m)}\right\} \quad (1)$$

where Y is the set of possible brightness values and $p(y|I, c)$ is the probability density function of pixel value y included in image I under lighting condition c. The lighting condition c could refer to either normal, excessively dark, excessively bright, or nonuniform lighting. The Bhattacharyya distance could be either a complete match (value of 0 (minimum)) or a complete mismatch (value of 1 (maximum)). Two examples of the Bhattacharyya distances between two randomly selected images are shown in Figure 2. As shown, there was a value close to 1 between two images with large differences, while there was a relatively small value between similar images. The averages of the Bhattacharyya distances for each image modality (optical, depth, and IR images) are presented in Table 1, where the results for the three combinations of lighting conditions (normal–excessively dark, normal–nonuniform and normal–excessively bright) are given for each vowel. As expected, the average distances of the optical images was approximately one for all vowels, whereas smaller distances were observed with the IR images. For the depth image, the results approximated the middle of the two types of images. The significance test also showed that there were remarkable differences in the Bhattacharyya distances between an optical image and other images ($p = 7.75e^{-10}$ and $p = 4.5e^{-14}$ for optical–IR and optical–depth, respectively). This indicates that the brightness distribution of the optical images varied greatly depending on the ambient light. On the other hand, even large changes in lighting conditions only caused minor changes in the IR and depth images. For the pair of normal and nonuniform lighting conditions, the average distances for optical images were slightly decreased, but remained significantly higher than those for the IR images ($p = 4.67e^{-8}$). Similar results were commonly observed for all vowels.

In conclusion, it appears that the variation in the captured images when using ambient light was insignificant for the IR and depth images. Such properties of the IR/depth images are expected to be helpful in implementing a robust automatic lip-reading scheme against variations in lighting conditions.

Table 1. Bhattacharyya distances of optical, depth, and IR images for five vowels ("/a:/", "/i:/", "/u:/", "/e:/", and "/o:/") according to the different lighting conditions.

Condition Pair	Image	Vowel				
		/a:/	/i:/	/u:/	/e:/	/o:/
Normal-to-excessively dark	Optical	0.650	0.678	0.662	0.666	0.662
	Depth	0.076	0.082	0.063	0.076	0.055
	IR	0.087	0.084	0.083	0.084	0.090
Normal-to-nonuniform	Optical	0.445	0.467	0.443	0.460	0.452
	Depth	0.194	0.234	0.208	0.202	0.192
	IR	0.095	0.096	0.092	0.091	0.096
Normal-to-excessively bright	Optical	0.608	0.612	0.553	0.604	0.566
	Depth	0.161	0.208	0.184	0.188	0.171
	IR	0.084	0.084	0.089	0.087	0.095

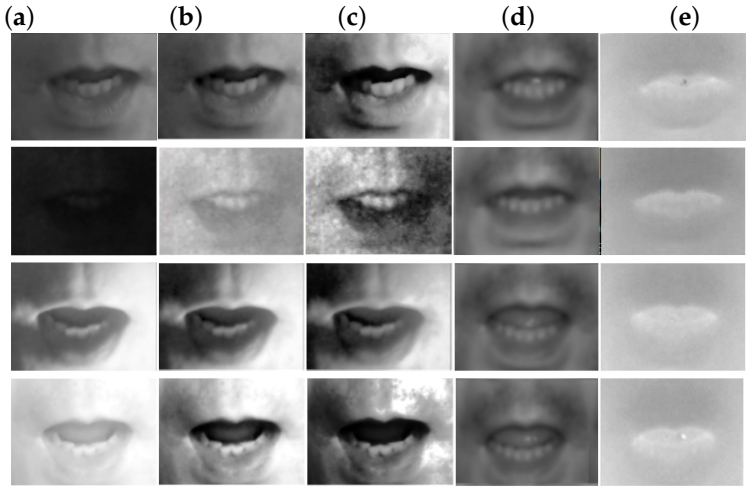

Figure 1. Examples of the captured images for voiced "/a:/" under normal lighting (1st row), dark lighting (2nd row), nonuniform illumination (3rd row), and bright lighting (4th row). Each column corresponds to (**a**) unprocessed optical images, (**b**) intensity-normalized optical images, (**c**) histogram-equalized optical images, (**d**) IR images, and (**e**) depth images.

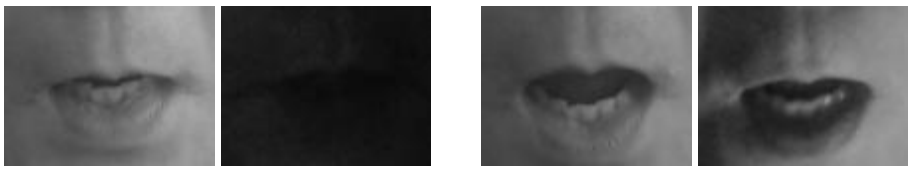

Figure 2. Examples of the Bhattacharyya distances between the two optical images; each figure presented below represents the Bhattacharyya distance between the two images above.

3. Implementation of Visual Speech Recognition

It is well known that a sequence of speech signals can be modeled by means of a quasi-stationary random process. Accordingly, the problems of visual speech recognition can be formulated as a time sequence classification. Due to the advances in deep neural networks (DNNs), recently developed visual speech recognition systems were implemented using sequence-processing networks, such as long short-term memory networks (LSTMs), gated recurrent units (GRUs), attention-based transformers, and temporal convolutional networks (TCNs).

Although the usage of a convolutional neural network (CNN) has the advantage of being able to use raw entire images directly as a neural network input, a face localization step should be taken into consideration to improve the recognition accuracy. To this end, some of the facial landmarks were located to extract only the speaker's lips as the region of interest (ROI) and as a feature input for the visual front-end [52]. This could be regarded as another type of recognition process, and performance degradation could occur depending on environmental factors, such as ambient light. In this study, errors in face localization due to lighting conditions were not considered. Hence, ground-truth facial landmarks were assumed to be a given. The set of features for visual speech recognition is given by

$$S_F = \{|F(u,v,w)||\text{if } M(u,v,w) = 1\} \tag{2}$$

where $|\bar{F}(u,v,w)|$ is the magnitude of 3D-DFT for the lip image $f(x,y,t)$:

$$F(u,v,w) = \sum_{x=0}^{N_x-1} \sum_{y=0}^{N_y-1} \sum_{t=0}^{N_t-1} f(x,y,t) e^{-j2\pi(\frac{ux}{N_x} + \frac{vy}{N_y} + \frac{wt}{N_t})} \tag{3}$$

where $N_x \times N_y$ is the size of the lip image, which was 106×80 and 90×70 for the optical images and the IR/depth images, respectively. N_t is the maximum length of the video, which was chosen as 47 in this study. The zero-frames (all pixel values are zero) were padded before/after the image sequence when the length of the input video was shorter than N_t, whereas a number of first/last frames were removed when the video length was longer than N_t. The mask matrix $M(u, v, w)$ was built as follows:

$$M(u,v,w) = \begin{cases} 1 & \text{if } |\bar{F}(u,v,w)| \geq F_{Th} \\ 0 & \text{otherwise} \end{cases}$$

where $|\bar{F}(u, v, w)|$ is the average of the FT magnitudes, as computed from the training dataset. The threshold of F_{Th} was determined so that the number of the elements in S_F was equal to N_F (= the number of DFT magnitude coefficients). In the subsequent section, the classification accuracy is shown according to N_F. Since the 3D-DFT magnitude spectrum is not changed by spatial shifts, the performance degradation caused by lip displacement was partially alleviated. Moreover, the classification accuracy was affected neither by the variations in onset nor by the termination times of vocalization because the 3D-DFT magnitude spectrum was given freely from the video delay. High levels of redundancy were often observed in the lip images. This indicates that a more compact feature for speech recognition could be obtained by truncating the DFT spectrum.

The right side of Figure 3 displays a block diagram of the classification step. The structure of the DNN adopted in the present study took the form of a multi-layer perceptron (MLP). The 3D-DFT magnitude coefficients selected by the mask matrix were arranged in a 1D array, which was then inputted into a DNN. According to the experimental results, the best performance was obtained when the DNN contained three hidden layers and the numbers of the nodes in the hidden layers were set to $[0.75 \times N_i]$, $[0.5 \times N_i]$, and $[0.5 \times N_i]$, where $[N]$ is the nearest integer value of N, and N_i is the number of input nodes, which was equal to N_F. For all layers, the sigmoid activation function was adopted. The momentum constant α of the sigmoid active function was set to 0.7. Such an MLP configuration was determined using a validation dataset that was 10% of the entire learning dataset. The objective function was given by the cross-entropy between the actual outputs in the top layer and the targets.

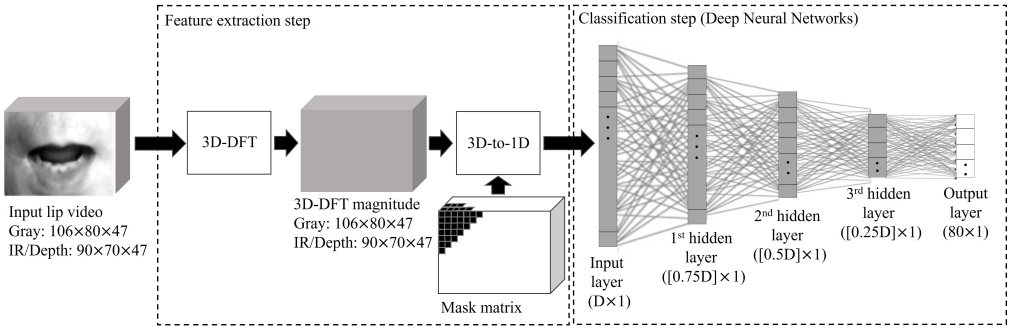

Figure 3. Block diagram of the proposed lip-reading system that includes the steps for feature extraction and classification steps.

4. Prediction of IR/Depth Images Using an Optical RGB Image

While IR/depth images have advantages over optical RGB images in various lighting environments, there are some issues to be addressed from an implementation perspective. Compared with RGB cameras that use natural light, the IR and depth cameras require a separate light source that emits the light of a wavelength appropriate to the IR range. This means that a separate space and drive circuit are necessary for the IR light source. NIR images can be acquired by RGB cameras with silicon sensors, and thus, it is possible to acquire RGB and IR images with one camera. However, the simultaneous acquisition of RGB and IR images is not possible and requires separate acquisition times. This leads to mismatches

between the two images due to an acquisition time difference. A special sensor is required when using a TOF camera to detect the phase differences in the light. If IR and depth images can be estimated from an RGB image, a lip-reading system with a high recognition rate can be implemented by using existing RGB cameras without adding/changing any hardware. Previous studies demonstrated the feasibility of using RGB images to predict a different domain for its application-specific representation [44]. Accordingly, in this paper, we propose a method to improve the accuracy of lip-reading by using IR/depth images estimated from RGB images instead of the actually captured IR/depth images.

Estimating IR/depth images from RGB images can basically be formulated as a problem of finding the pixel-by-pixel mapping rules between the two images. The underlying assumption is that a large amount of low-level information, such as the location of edges, is shared between the two images [49]. Despite the shared information between the two images, each image has unique characteristics that cannot be explained by simple dependency relationships. Therefore, the correspondence between the two images was primarily represented by nonlinear mapping rules, such as deep neural networks [38–40,42–45,49,53]. Typically, a CNN was adopted to estimate the IR/depth images from optical RGB images. The architecture of the CNN used in this study was similar to that of U-net [53], as illustrated in Figure 4. In a previous study, a dual encoder–decoder-based architecture with different depths [44] and conditional generative adversarial networks [45,49] was employed to estimate IR images from optical RGB images. We performed evaluation tests on these two forms of architecture from a recognition accuracy perspective. No clear performance advantage over the structure shown in Figure 3 was observed in our experiments. At each layer, the convolution kernel size, image depth, and pooling type were different from those used in U-net; these values were empirically determined using the validation dataset.

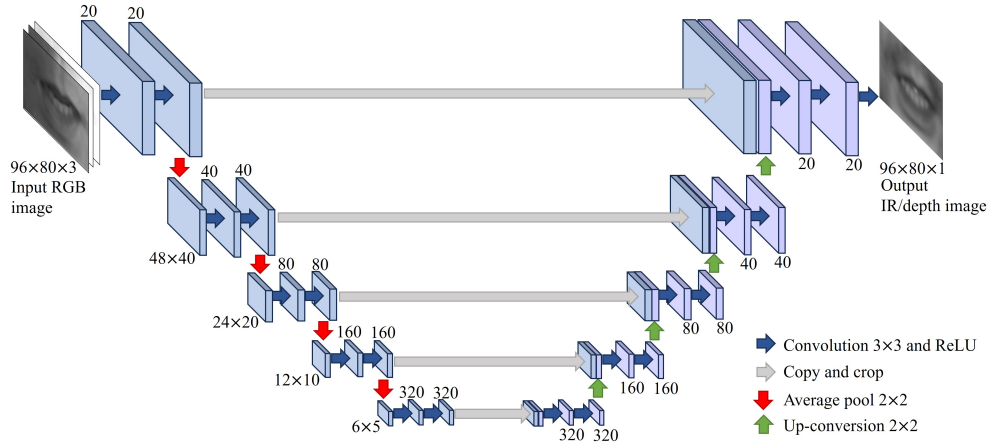

Figure 4. The architecture of the CNN that transformed RGB images to depth (or IR) images.

A backpropagation algorithm that involved the minimum mean square error (MMSE) criterion was used to train the CNN. The objective function was given by the mean square error between the estimated and actual IR (or depth) images as follows:

$$E = \frac{1}{N}\sum_{n=1}^{N}\left\{\mathcal{F}(\mathbf{W},\mathbf{X}_n) - \mathbf{Y}_n\right\}^2 \tag{4}$$

where $\mathcal{F}(\mathbf{W},\mathbf{X}_n)$ is the output of the CNN with the set of kernels $\mathbf{W}$, and the input RGB image $\mathbf{X}_n$ is given. $\mathbf{Y}_n$ denotes the target image (IR or depth image) at frame index n and N is the total number of training images. A stochastic gradient descent algorithm was performed in mini-batches with multiple epochs to improve the learning convergence. The

updated estimate of the set of kernels **W** with a learning rate λ was computed iteratively as follows:

$$\mathbf{W}_{n+1} = \mathbf{W}_n - \lambda \bigtriangledown_{\mathbf{W}} E \tag{5}$$

The use of MMSE as a loss function seems to be an appropriate choice from the perspective of increasing the similarity between the estimated image and the true image. In this study, however, the performance needed to be evaluated in terms of recognition accuracy rather than by similarities between the two images. We assumed that the closer the estimated image was to the real image, the closer the performance of the speech recognition would be to that of using the real image.

5. Experimental Results

5.1. Data Preparation

The device shown in Figure 5 was constructed to simultaneously acquire optical RGB images, infrared images, and depth images. Since the TOF camera employed in this study was capable of simultaneously acquiring and storing depth and IR images, both an RGB camera (webcam) and TOF camera were used. In addition to the cameras, a microphone was fitted to record the subject's voice. This was used for synchronization between each video. The distance between the cameras and the subject's face was fixed at approximately 50 cm, and the lenses of the TOF camera and optical RGB camera were adjusted to focus on the same point. Although the lenses of both cameras were pointed at the same point, a certain degree of discrepancy was observed between the two images captured by each camera, as shown in Figure 6. Such a mismatch could have been caused by a poor IR/depth image estimation from the RGB images. To partially resolve this issue, markers were placed on the subjects' faces prior to acquisition. The colors and shapes of the markers were chosen to be clearly visible in both the optical RGB and IR images, as shown by the black dots on green backgrounds in Figure 6. Note that the markers were placed a certain distance away from the lips to prevent them from being visible in the lip image.

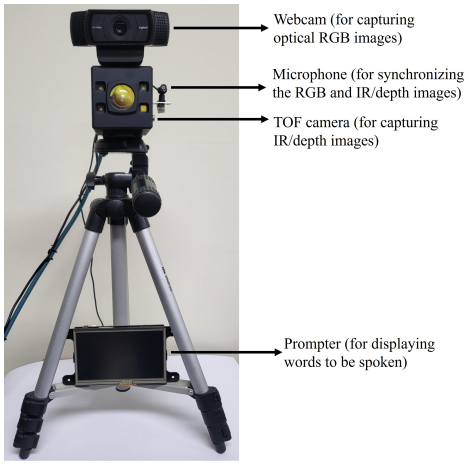

Figure 5. A photograph of the device used for capturing images.



Figure 6. An example of the captured face images: (**left**) optical RGB image and (**right**) IR image.

The two images were matched using an affine transformation on a pixel-by-pixel basis. Since the IR images were estimated using the optical RGB images, affine transformations were carried out on the RGB images. The reference points required for the affine transformation were the center points of each marker, which were obtained manually through visual observation. The accuracy of matching in the affine transformations is known to improve as the number of reference points increases. Hence, it is desirable to attach more markers to the face in order to minimize the geometric mismatches between the optical and IR images. Our experimental results show that four markers around the lips was sufficient to obtain well-matched images under normal lighting conditions. However, in cases of uneven lighting or excessively dark lighting, certain markers became invisible, and thus, more markers were required in multiple facial locations. As a result, using a total of seven markers, including the midpoint between the eyebrows and two points below the eyes, seemed to be a good compromise between subject discomfort and matching accuracy, even under abnormal lighting conditions.

A video dataset was obtained from three healthy subjects (two males and one female, with ages ranging from 23 to 55 years) without known articulatory problems. The dataset consisted of 80 phonetically balanced words in the Korean language. Each subject repeatedly uttered each word 100 times under normal lighting conditions. Data augmentation was achieved by asking the subjects to speak the same words in styles that would appear as different as possible. The ratio of training, testing, and validation of the datasets was 10:2:1. Additional audio-visual data were acquired 10 times for each word under four different lighting conditions (dark, bright, and two nonuniform). The additional data were gathered by adjusting the current of the LED illumination and controlling the LED direction. During the recording, subjects were asked to keep their heads as still as possible by placing the center of their mouth at a certain point. Landmark-based segmentation [52] followed by manual corrections were carried out on the captured images to extract the subject's mouth region. Each of the neural networks was individually trained using the visible images, IR images, and depth images acquired under normal lighting conditions. Then, each trained neural network was applied to each test dataset under normal lighting conditions and each dataset acquired under abnormal lighting conditions (excessively dark, excessively bright, and nonuniform lighting conditions).

5.2. Recognition Accuracy for Each of the Captured Images

The accuracies for each modality according to the lengths of input features are plotted in Figure 7. For comparison, the CNN-based recognition schemes (ConvNet+LSTM and 3D-ConvNet in [54]) were also employed, wherein raw optical images were used as input into the CNN. Note that since such networks were originally devised for action recognition [54], a small modification was made to the architecture of these networks. For all modalities, the recognition accuracy increased with an increase in the length of the input features. A maximum accuracy of 95.92% was obtained when using the 1000 DFT magnitude coefficients of an IR image, which was 6% higher than that of the optical image. Using the truncated 3D-DFT magnitude spectra of an optical image revealed a higher recognition rate compared with that of the CNN-based methods. This could have been due to the space- and time-delay-invariant properties of the 3D-DFT magnitude spectra. For all modalities, the recognition accuracy was almost saturated when the length of input features exceeded 1000.

Thus far, the discussion has considered results that were obtained from images acquired under normal lighting conditions. The recognition accuracies from images acquired under different lighting conditions are presented in Figure 8. The superiority of the nonvisible images is clear. A maximum accuracy of 85.5% was achieved using IR images when the length of the input features was 750. For optical images, however, the maximum accuracy was 17.0%. This increased to 24.4% when the intensity-normalized images were adopted, which indicates that only limited improvements could be achieved by globally changing the statistical properties of a given image. Even when ConvNet+LSTM and 3D-ConvNet

were applied to the optical images acquired under abnormal lighting conditions, the performance remained significantly lower than when using IR images. Considering that the variation in the captured IR images caused by ambient light was very small, such results were somewhat expected. A slight decrease in the recognition rate was observed for the IR images compared with the images captured under normal lighting conditions. This was assumed to have been caused by a small difference in the IR images due to changes in the lighting conditions.

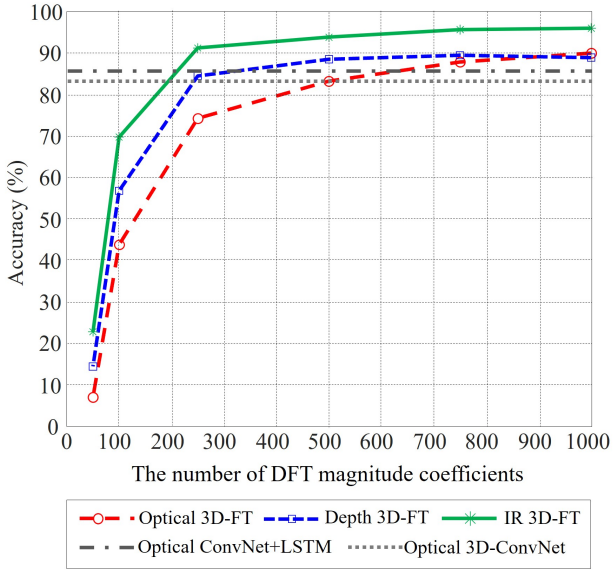

Figure 7. Recognition accuracies for different recognizers according to the number of DFT magnitude coefficients under normal lighting conditions.

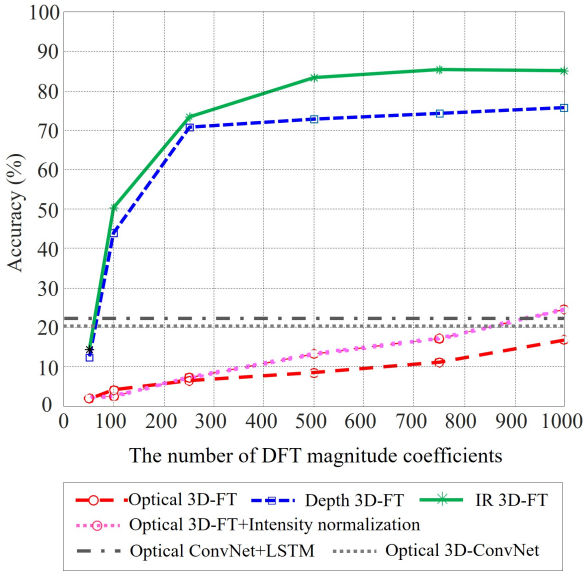

Figure 8. Recognition accuracies for the different recognizers according to the number of DFT magnitude coefficients under abnormal lighting conditions.

The recognition accuracy of the depth images acquired under abnormal lighting conditions approximated that of the IR image when using a relatively short input feature (<500). However, there was a clear difference in the recognition accuracy when a longer (≥ 500) input feature was employed. The pixel values of the depth image may be changed by even a small front/rear displacement of the subject, which cannot be mitigated by

adopting a 3D-DFT magnitude spectrum. A recognition rate that was lower than that of the IR image was partially due to such variations.

5.3. Image Conversion Results

The main objective of the image conversion (RGB-to-IR and RGB-to-depth) was not only to evaluate the estimation accuracy, but also to observe whether the estimated IR/depth images showed suppressed fluctuations like the actual acquired images, even under different lighting conditions. A total of 228,613 images were used to build the depth/IR image estimation rules from RGB images, and the evaluation was carried out for 8270 images. The image estimation was performed only on the subject's mouth region, not on the entire image, and the size was 96 × 80 pixels (W × H). The images used for the construction of the estimation rules and evaluation included images acquired under a variety of lighting conditions (i.e., normal, excessively dark, nonuniform, and excessively bright, as shown in the first row of Figure 9).

As objective measures for evaluating the estimation accuracy, the peak signal-to-noise ratio (PSNR), mean squared error (MSE), and structural similarity index map (SSIM) [55] were employed in this study. The results are presented in Tables 2 and 3. For the RGB-to-IR conversion, the results for six different neural networks are shown to compare the performances according to the architecture of the neural network. The hyperparameters used to change the architecture were the pooling type, convolution kernel size, and the depth of the intermediate images at each layer. Note that for each neural network, the depth of the intermediate images at each layer was heuristically determined by increasing or decreasing the values adopted in U-net.

The experimental results show that the estimation accuracy was not significantly affected by the architecture of the neural network. However, the performance was slightly degraded when the kernel size was relatively large (7 × 7). This was likely due to the relatively small size of the images used and the fact that the images were estimated without sufficiently reflecting the detailed structure of the input RGB image. Compared with the IR images, the depth images revealed a very high level of estimation accuracy. This was because the resolution of the employed depth camera was not sufficient to show the fine curvature of the facial surface, and thus, the details commonly observed in the optical and IR images did not appear in the depth images. This means that the RGB-to-depth image conversion was a simpler problem than the RGB-to-IR image conversion, which resulted in a much higher level of estimation accuracy for the depth image estimation.

Table 2. Performance of RGB-to-IR image estimation for each network architecture.

Pooling Type	Kernel Size	Image Depth at Each Layer	PSNR (dB)	MAE	SSIM
Average	5 × 5	20-40-80-160-320	30.20	6.90	0.886
Average	5 × 5	24-48-96-192-384	30.55	6.57	0.890
Average	7 × 7	24-48-96-192-384	29.78	7.56	0.876
Average	3 × 3	20-40-80-160-320	30.00	6.98	0.882
Max	5 × 5	20-40-80-160-320	30.15	6.92	0.885
Average	5 × 5	16-32-64-128-256	30.17	6.79	0.885

Table 3. Performance of RGB-to-depth image estimation for each network architecture.

Pooling Type	Kernel Size	Image Depth at Each Layer	PSNR (dB)	MAE	SSIM
Average	5 × 5	20-40-80-160-320	46.13	1.23	0.984
Average	5 × 5	24-48-96-192-384	45.90	1.21	0.992

An example of an IR/depth image estimated from an RGB image is shown in Figure 9, along with the actual captured images for comparison. As shown in the figure, the estimated

images (third and fifth rows) were in close visual agreement with the actual images acquired by the camera. It is noteworthy that even though the optical image showed a great deal of fluctuation depending on the lighting conditions, the estimated IR/depth image showed almost no such fluctuation. This indicates that converting an RGB image to a depth/IR image could be used not only to obtain each image itself but also to suppress large variations in the image due to different lighting conditions. This was also confirmed by comparing the Bhattacharyya distances between the images acquired under normal and abnormal lighting conditions, as shown in Table 4. The distances shown in the table were measured using actual camera acquisitions for optical images, but the images estimated from RGB images were used in the cases of IR and depth images. Similar to the actual acquired IR/depth images, the estimated images showed less variation for the same vowel utterance, even when the lighting conditions changed significantly. The results in Table 1 show the Bhattacharyya distances for the actual IR/depth images, but the estimated images revealed slightly higher distance values. This indicates that the image conversion did not reduce the image fluctuations due to lighting changes as much as the actual image. Nevertheless, the residual fluctuations seen in the estimated images were not visibly significant, as shown in Figure 9.

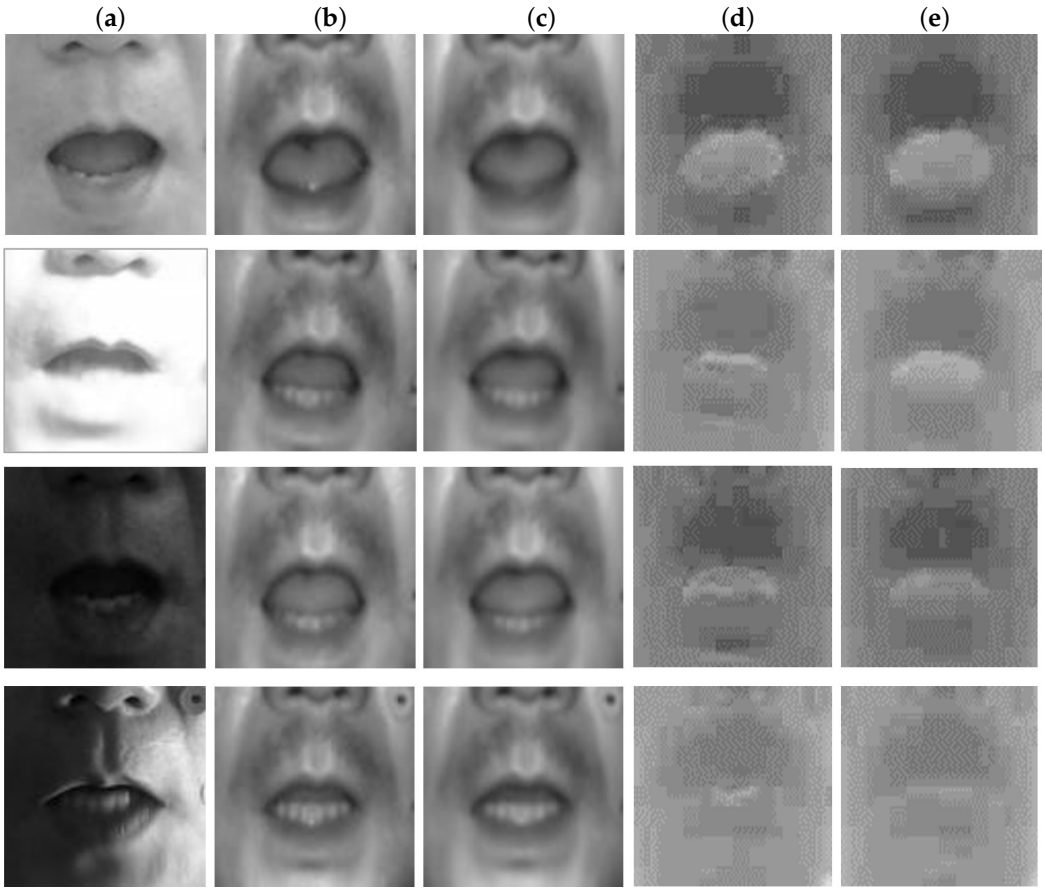

Figure 9. Examples of the captured (real) images and estimated images for voiced "/a:/" under normal (1st row), excessively bright (2nd row), excessively dark (3rd row), and nonuniform illumination (4th row). Each column corresponds to (**a**) optical images, (**b**) real IR images, (**c**) estimated IR images from corresponding optical images, (**d**) real depth images, and (**e**) estimated depth images from corresponding optical images. Notes: All depth images were adjusted for contrast/brightness to enhance the visibility.

In conclusion, the proposed image conversion framework was shown to be effective in suppressing various fluctuations in the image, which is expected to improve the performance of speech recognition under various lighting conditions.

Table 4. Bhattacharyya distances of optical, estimated depth, and estimated IR images according to the different lighting conditions for five vowels "/a:/", "/i:/", "/u:/", "/e:/", and "/o:/".

Condition Pair	Image	Vowel				
		/a:/	/i:/	/u:/	/e:/	/o:/
Normal-to-excessively dark	Optical	0.650	0.678	0.662	0.666	0.662
	Depth	0.144	0.124	0.124	0.128	0.121
	IR	0.269	0.259	0.270	0.257	0.275
Normal-to-nonuniform	Optical	0.445	0.467	0.443	0.460	0.452
	Depth	0.211	0.230	0.219	0.217	0.199
	IR	0.273	0.269	0.279	0.260	0.280
Normal-to-excessively bright	Optical	0.608	0.612	0.553	0.604	0.566
	Depth	0.204	0.226	0.227	0.234	0.208
	IR	0.262	0.255	0.269	0.252	0.271

5.4. Recognition Accuracy for Robust Automatic Lip-Reading against Variations in Lighting Condition

With a sufficient amount of (optical/IR or depth) images acquired under a variety of lighting conditions, the following methods were considered automatic lip-reading techniques that were robust for producing lighting variations:

(1) Baseline method: The optical images acquired under various lighting conditions were divided into training and evaluation datasets, and the classification rules for lip-reading were built from the training data, where these rules were evaluated using the evaluation dataset.
(2) Conversion-based post-processing approach: Classification rules for automatic lip-reading were built from the true IR (or depth) images acquired under various lighting conditions. Image conversion rules (RGB-to-IR or RGB-to-depth) were constructed using an optical (IR or depth) image pair dataset. The converted (IR or depth) image from the given optical image was used for the evaluation.
(3) Using classification rules built by the converted images: Image conversion rules were built in the same way as in (2). All the training data, including images acquired under various lighting conditions, were converted using the conversion rules. Then, the classification rules were built using the converted images. In the evaluation, the optical RGB image was first converted to an IR (or depth) image, and recognition was carried out on the converted image using the constructed classification rules.

The block diagrams of each method are presented in Figure 10. Note that the set of IR or depth images was used only in the training stage, while only optical images were used in the evaluation. Therefore, any devices for acquiring IR/depth images were not required in the evaluation.

The performance of each method in terms of the classification accuracy (%) is presented in Figure 11. The accuracies for each method were plotted according to the length of the truncated 3D-DFT magnitude spectra. For small lengths ($\leq$700), the baseline method without IR images showed relatively low accuracy compared with the method using IR images (methods (2) and (3)). This indicated that the conversion of RGB images to IR images improved the accuracy to some extent when an FT spectra with a relatively high degree of approximation was used as the input feature. As the length of the input feature increased, the baseline method and the conversion-based post-processing technique resulted in a similar performance.

The highest accuracy for the baseline method was 86.62%, which was slightly lower than that of the optical images acquired under normal lighting ($\approx$90%). Considering that the evaluation dataset used in this experiment included images acquired under abnormal lighting conditions, as well as normal lighting, it appears that the recognition rate was significantly improved compared with using a classification rule trained without consid-

ering lighting conditions. When the classification rules trained on the converted images were used, the recognition rate was always high compared with the other two methods for all feature variable lengths. The highest accuracy of 89.35% was also achieved when using a classification rule learned from the converted image and a length of 1600 input features. This was almost identical to the highest accuracy value obtained from the optical images acquired under normal lighting. Such a result suggests that converting all optical images to IR images (regardless of the lighting conditions) and then generating classification rules from the converted images will improve the recognition accuracy. However, neither of the methods using IR images estimated from RGB images exceeded the highest recognition rate when using true IR images acquired under normal lighting conditions (about 96%). This could be inferred from the results of previous experiments, where the IR image estimated from an optical RGB image showed a slight difference from an actual IR image. The difference in recognition rate between the actual IR images and estimated IR images is expected to be reduced by improved image conversion techniques. Note that no significant improvement in recognition rate was observed when the length of the input variables exceeded 2000.

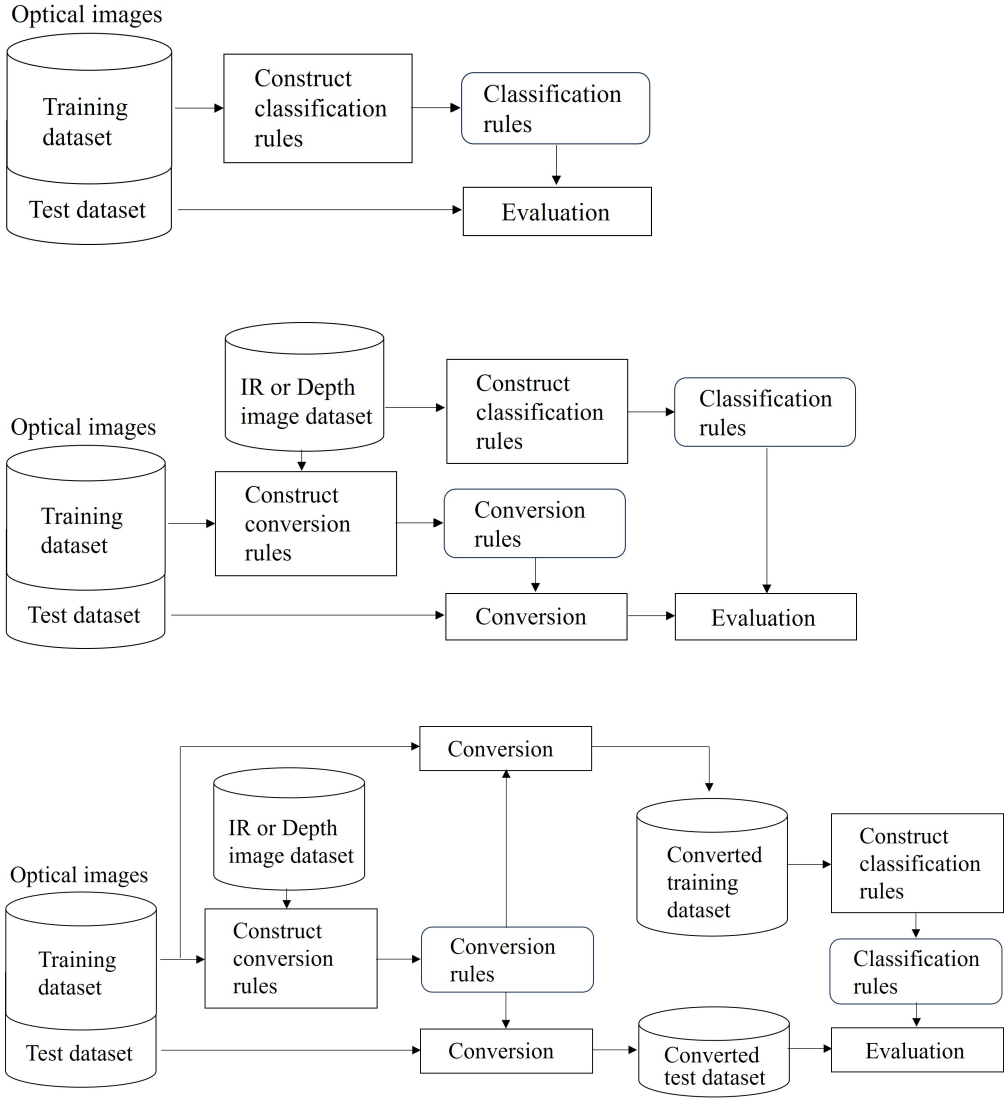

Figure 10. The block diagrams of robust automatic lip-reading against variations in lighting conditions. **Top**: baseline method. **Middle**: conversion-based post-processing approach. **Bottom**: using classification rules built by the converted image.

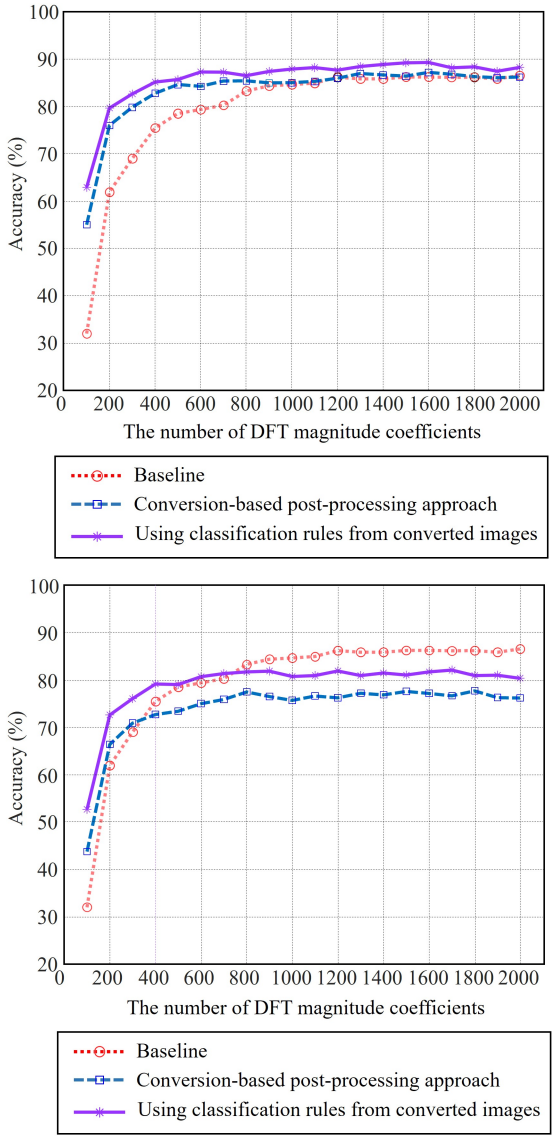

Figure 11. Recognition accuracies for the different recognizers according to the number of DFT magnitude coefficients. **Top**: using RGB-to-IR image conversion. **Bottom**: using RGB-to-depth image conversion.

When using the converted depth image, the recognition rate was slightly lower overall compared with using the converted IR image (bottom of Figure 11). Similar to the case when using the IR image, the classification rules trained on the converted image revealed a higher accuracy when using a very short length of input features (<400). A relatively high accuracy was achieved by using a classification rule trained on the mixed images acquired under normal and abnormal lighting conditions when the length of the input feature exceeded 700. For the depth images, using both methods of conversion as post-processing and generating classification rules from the converted images resulted in lower classification accuracies compared with using classification rules obtained from the mixed dataset. This was somewhat unexpected given that the RGB-to-depth image conversion revealed a significantly higher conversion performance than the RGB-to-IR conversion. There are two possible explanations for these conflicting results: (1) Even when using images acquired with a real camera, the depth images showed somewhat lower recognition rates than the IR images. (2) Even if the image conversion (RGB-to-depth) error was small, the distortion caused by the conversion had a significant impact on lip-reading. From such

experimental results, it can be said that in terms of the lip-reading accuracy, depth images were not useful compared with the use of optical images alone.

5.5. Comparison with Existing Contrast Correction Methods

To evaluate the usefulness of the proposed image correction techniques using IR/depth images, a comparison with existing contrast correction methods was performed. The methods selected for comparison included extended exposure fusion (EEF) [21], homomorphic filtering (HF) [22], and contrast-limited adaptive histogram equalization (CLAHE) [24]. Note that although the comparison methods were primarily developed for visual enhancement, the performance evaluation was done from the perspective of the accuracy of speech recognition, not from the visual perspective of the processed images. For the same evaluation dataset used in the previous section, the speech recognition rates for the images processed by each method were compared. The third method explained in the previous section, namely, using classification rules built by the converted images, was employed as the speech recognition method, in which the length of the truncated 3D-DFT magnitude spectra was fixed to 2000.

The results are summarized in Table 5, where "regIR" refers to the method proposed in this study that used IR images converted from RGB images. The results show that the method using converted IR images had a significantly higher recognition rate compared with the other methods. The superiority of the conversion-based method was mainly due to its significantly high recognition rates for the captured images under nonuniform lighting. Since the conversion-based method produces IR images that are less influenced by ambient visible light sources, such a method can be regarded as similar to the homomorphic filtering methods, which suppress the illumination components and relatively emphasize the reflective components. For the images captured under nonuniform lighting, the illumination components were spatially varied. Accordingly, a simple high-pass filtering is not as effective as suppressing light fluctuations as it is for IR or depth images. This can be one of the reasons for the relative poor performance of homomorphic filtering. Among the comparison methods, the CLAHE was the only one that took into account the spatial characteristics of the image. For this reason, the recognition rate was somewhat higher than other comparison methods. The EEF method also processed the images globally, which resulted in poor recognition rates for images acquired under nonuniform lighting.

Table 5. Recognition accuracies for the different contrast correction methods.

Methods	**EEF**	**CLAHE**	**HF**	**regIR**
Acc. (%)	76.35	80.28	73.44	88.33

Consequently, such results indicate that automatic lip-reading techniques using IR/depth images predicted from RGB images are effective in improving accuracy compared with previous contrast enhancement methods, despite the cost of acquiring additional IR/depth images. The results also suggest the feasibility of a new form of contrast enhancement algorithm with regression analysis. Although the IR and visible images were not exact matches, they were visually similar and can be thought of as highly correlated. Therefore, IR images can be chosen as a good reference for generating correction rules.

6. Conclusions

Recognition rate degradation due to image variations is observed in automatic lip-reading. In this study, IR images and depth images were employed to partially overcome the problem of the decreased recognition rate due to fluctuations, particularly in lighting conditions. The effectiveness of IR/depth images was first compared with that of optical images and investigated from an image-based speech recognition perspective. The superiority of the IR/depth images under abnormal lighting conditions was clearly demonstrated by the experimental results. IR images can be acquired by using currently available RGB cameras with small modifications. Hence, it is easy to implement automatic lip-reading

with a high recognition rate, even under abnormal lighting conditions. Acquiring IR images, however, requires special lighting and optical filters to block visible light and emit only infrared light. This means that hardware modifications are required to implement IR- or depth-image-based lip-reading systems on mobile devices with existing RGB cameras.

In this study, an image conversion technique was applied to maintain the advantages of IR/depth-image-based lip-reading using devices equipped with existing RGB cameras without hardware changes. To obtain IR or depth images from RGB images, we modified the existing CNNs that were previously used for image-to-image conversion. The main purpose of using the converted IR/depth images from RGB images was to achieve a higher rate of speech recognition from IR/depth images while suppressing fluctuations in the lighting conditions, even if a converted image was used instead of the true image. In other words, image conversion was used to suppress the fluctuations in ambient light, not to obtain actual information from the infrared spectrum. According to the experimental results, the highest recognition rate was achieved in cases when the classification rules were constructed using IP images converted from RGB images. In this case, the maximum recognition rate was almost identical to that obtained from optical images acquired under normal lighting conditions. The adopted methods in this study were not limited to image-based speech recognition and could also be applied to other types of image recognition that suffer from poor recognition rates due to variations in lighting conditions. Future research will be conducted in this direction.

Funding: This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. 2022R1F1A10689-791220682073250102).

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J.M.; Brumberg, J.S. Silent speech interfaces. *Speech Commun.* **2010**, *52*, 270–287. [CrossRef]
2. Fernandez-Lopez, A.; Sukno, F.M. Survey on automatic lip-reading in the era of deep learning. *Image Vis. Comput.* **2018**, *78*, 53–72. [CrossRef]
3. Fenghour, S.; Chen, D.; Guo, K.; Li, B.; Xiao, P. Deep Learning-Based Automated Lip-Reading: A Survey. *IEEE Access* **2021**, *9*, 121184–121205. [CrossRef]
4. Vanegas, O.; Tokuda, K.; Kitamura, T. Location normalization of HMM-based lip-reading: Experiments for the M2 VTS database. In Proceedings of the International Conference on Image Processing, Kobe, Japan, 24–28 October 1999; pp. 343–347.
5. Movellan, J.R. Visual speech recognition with stochastic networks. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 28 November–1 December 1994; pp. 851–858.
6. Messer, K.; Matas, J.; Kittler, J.; Luettin, J.; Maitre, G. XM2VTSDB: The extended M2VTS database. In Proceedings of the International Conference on Audio Video-Based Biometric Person Authentication, Washington, DC, USA, 22–24 March 1999; pp. 965–966.
7. Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, USA, 28 June–2 July 2011; pp. 1–8.
8. Ortega, A.; Sukno, F.; Lleida, E.; Frangi, A.F.; Miguel, A.; Buera, L.; Zacur, E. AV@CAR: A Spanish multichannel multimodal corpus for in-vehicle automatic audio-visual speech recognition. In Proceedings of the International Conference on Language Resources and Evaluation, Lisbon, Portugal, 26–28 May 2004; pp. 1–4.
9. Matthews, I.; Cootes, T.F.; Bangham, J.A.; Cox, S.; Harvey, R. Extraction of visual features for lipreading. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 198–213. [CrossRef]
10. Yanjun, X.; Limin, D.; Guoqiang, L.; Xin, Z.; Zhi, Z. Chinese audiovisual bimodal speeeh database CAVSR1.0. *Acta-Acust.* **2000**, *25*, 42–44.
11. Kumar, K.; Chen, T.; Stern, R.M. Profile view lip reading. In Proceedings of the IEEE International Conference on Acoustic Speech Signal Processing, Honolulu, HI, USA, 16–20 April 2007; pp. IV429–IV432.
12. Mesbah, A.; Berrahou, A.; Hammouchi, H.; Berbia, H.; Qjidaa, H.; Daoudi, M. Lip reading with Hahn convolutional neural networks. *Image Vis. Comput.* **2019**, *88*, 76–83. [CrossRef]
13. Ma, P.; Martinez, B.; Petridis, S.; Pantic, M. Towards Practical Lipreading with Distilled and Efficient Models. In Proceedings of the IEEE International Conference on Acoustic Speech Signal Processing, Toronto, ON, Canada, 6–11 June 2021; pp. 7608–7612.

14. Fenghour, S.; Chen, D.; Guo, K.; Xiao, P. Lip reading sentences using deep learning with only visual cues. *IEEE Access* **2020**, *8*, 215516–215530. [CrossRef]
15. Assael, Y.; Shillingford, B.; Whiteson, S.; Freitas, N.D. Lipnet: End-to-end sentence-level lipreading. *arXiv* **2016**, *preprint*. arXiv:1611.01599.
16. Chung, J.S.; Senior, A.; Vinyals, O.; Zisserman, A. Lip reading sentences in the wild. In Proceedings of theComputer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3444–3453.
17. Tan, X.; Triggs, B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Process.* **2010**, *19*, 1635–1650.
18. Kalaiselvi, P.; Nithya, S. Face Recognition System under Varying Lighting Conditions. *IOSR J. Comput. Eng.* **2013**, *14*, 79–88. [CrossRef]
19. Zhu, J.-Y.; Zheng, W.-S.; Lu, F.; Lai, J.-H. Illumination invariant single face image recognition under heterogeneous lighting condition. *Pattern Recognit.* **2017**, *66*, 313–327. [CrossRef]
20. Jacobsen, N.; Deistung, A.; Timmann, D.; Goericke, S.L.; Reichen-bach, J.R.; Gullmar, D. Analysis of Intensity Normalization for Optimal Segmentation Performance of a Fully Convolutional Neural Network. *Z. Fur Med. Phys.* **2019**, *29*, 128–138. [CrossRef]
21. Hessel, C.; Morel, J.-M. An extended exposure fusion and its application to single image contrast enhancement. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 137–146.
22. Chavarín, Á.; Cuevas, E.; Avalos, O.; Gálvez, J.; Pérez-Cisneros, M. Contrast enhancement in images by homomorphic filtering and cluster-chaotic optimization. *IEEE Access* **2023**, *11*, 73803–73822. [CrossRef]
23. Lee, P.-H.; Wu, S.-W.; Hung, Y.-P. Illumination compensation using oriented local histogram equalization and its application to face recognition. *IEEE Trans. Image Process.* **2012**, *21*, 4280–4289. [CrossRef] [PubMed]
24. Suharyanto; Hasibuan, Z.A.; Andono, P.N.; Pujiono, D.; Setiadi, R.I.M. Contrast limited adaptive histogram equalization for underwater image matching optimization use SURF. *J. Phys. Conf. Ser.* **2021**, *1803*, 012008. [CrossRef]
25. Zheng, M.; Qi, G.; Zhu, Z.; Li, Y.; Wei, H.; Liu, Y. Image Dehazing by an Artificial Image Fusion Method Based on Adaptive Structure Decomposition. *IEEE Sens. J.* **2020**, *20*, 8062–8072. [CrossRef]
26. Zhu, Z.; Wei, H.; Hu, G.; Li, Y.; Qi, G.; Mazur, N. A Novel Fast Single Image Dehazing Algorithm Based on Artificial Multiexposure Image Fusion. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 5001523. [CrossRef]
27. Sugimura, D.; Mikami, T.; Yamashita, H.; Hamamoto, T. Enhancing color images of extremely low light scenes based on RGB/NIR images acquisition with different exposure times. *IEEE Trans. Image Process.* **2015**, *24*, 3586–3597. [CrossRef]
28. Salamati, N.; Fredembach, C.; Susstrunk, S. Material classification using color and nir images. In Proceedings of the IS&T/SID 17th Color Imaging Conference, Albuquerque, NM, USA, 9–13 November 2009.
29. Nonaka, Y.; Yoshida, D.; Kitamura, S.; Yokota, T.; Hasegawa, M.; Ootsu, K. Monocular color-IR imaging system applicable for various light environments. In Proceedings of the IEEE International Conference on Consumer Electronics, Las Vegas, NV, USA, 12–14 January 2018; pp. 1–5.
30. Shet, A.V.; Chinmay, B.S.; Shetty, A.A.; Shankar, T.; Hemavath, R.; Ramakanth, P. Face Detection and Recognition in Near Infra-Red Image. In Proceedings of the 6th International Conference on Computation System and Information Technology for Sustainable Solutions, Bangalore, India, 21–23 December 2022.
31. Nie, J.; Yan, L.; Wang, X.; Chen, J. A Novel 3D Convolutional Neural Network for Action Recognition in Infrared Videos. In Proceedings of the IEEE 4th International Conference on Information, Communication and Signal Processing, Shanghai, China, 24–26 September 2021; pp. 420–424.
32. Vishwakarma, D.K.; Grover, V. Hand gesture recognition in low-intensity environment using depth images. In Proceedings of the International Conference on Intelligent Sustainable Systems, Palladam, India, 7–8 December 2017; pp. 429–433.
33. Monno, Y.; Teranaka, H.; Yoshizaki, K.; Tanaka, M.; Okutomi, M. Single-sensor RGB-NIR imaging: High-quality system design and prototype implementation. *IEEE Sens. J.* **2019**, *19*, 497–507. [CrossRef]
34. Bhanu, B.; Han, J. Kinematic-based Human Motion Analysis in Infrared Sequences. In Proceedings of the 6th IEEE Workshop on Applications of Computer Vision, Orlando, FL, USA, 3–4 December 2002; pp. 208–212.
35. Wang, S.; Pan, B.; Chen, H.; Ji, Q. Thermal augmented expression recognition. *IEEE Trans. Cybern.* **2018**, *48*, 2203–2214. [CrossRef]
36. Torralba, A.; Oliva, A. Depth estimation from image structure. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 1226–1238. [CrossRef]
37. Herrera, J.L.; del-Blanco, C.R.; Garcia, N. Automatic depth extraction from 2D images using a cluster-based learning framework. *IEEE Trans. Image Process.* **2018**, *27*, 3288–3299. [CrossRef] [PubMed]
38. Li, B.; Shen, C.; Dai, Y.; Hengel, A.V.D.; He, M. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1119–1127.
39. Liu, F.; Shen, C.; Lin, G.; Reid, I. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 2024–2039. [CrossRef] [PubMed]
40. Cao, Y.; Wu, Z.; Shen, C. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 3174–3182. [CrossRef]
41. Wang, H.; Wang, Y.; Zhang, Q.; Xiang, S.; Pan, C. Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sens.* **2017**, *9*, 446. [CrossRef]

42. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic image segmentation with deep convolutional nets and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848. [CrossRef] [PubMed]
43. Song, M.; Lim, S.; Kim, W. Monocular depth estimation using Laplacian pyramid-based depth residuals. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 4381–4393. [CrossRef]
44. Aswatha, S.M.; Malladi, S.P.K.; Mukherjee, J. An encoder-decoder based deep architecture for visible to near infrared image transformation. In Proceedings of the 12th Indian Conference on Computer Vision Graphic and Image Processing, Jodhpur, India, 20–22 December 2021.
45. Yuan, X.; Tian, J.; Reinartz, P. Generating artificial near infrared spectral band from RGB image using conditional generative adversarial network. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* **2020**, *3*, 279–285. [CrossRef]
46. Aslahishahri, M.; Stanley, K.G.; Duddu, H.; Shirtliffe, S.; Vail, S.; Bett, K.; Pozniak, C.; Stavness, I. From RGB to NIR: Predicting of near infrared reflectance from visible spectrum aerial images of crops. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 1312–1322.
47. Shukla, A.; Upadhyay, A.; Sharma, M.; Chinnusamy, V.; Kumar, S. High-resolution NIR prediction from RGB images: Application to plant phenotyping. In Proceedings of the IEEE International Conference on Image Processing, Bordeaux, France, 16–19 October 2022; pp. 4058–4062.
48. Tang, C.; Zhang, H.; Kong, A.W.-K.; Craft, N. Visualizing vein patterns from color skin images based on image mapping for forensics analysis. In Proceedings of the 21st International Conference on Pattern Recognition, Tsukuba, Japan, 11–15 November 2012; pp. 2387–2390.
49. Keivanmarz, A.; Sharifzadeh, H.; Fleming, R. Vein pattern visualization using conditional generative adversarial networks. In Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, Auckland, New Zealand, 7–10 December 2020; pp. 1310–1316.
50. Sharma, N.; Hefeeda, M. Hyperspectral reconstruction from RGB images for vein visualization. In Proceedings of the 11th ACM Multimedia System Conference, Istandul, Turkey, 8–11 June 2020; pp. 77–87.
51. Kailath, T. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Comm. Technol.* **1967**, *15*, 52–60. [CrossRef]
52. Rizvi, Q.M. A review on face detection methods. *J. Manag. Dev. Inf. Technol.* **2011**, *11*, 1–11.
53. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, Toronto, ON, Canada, 5–9 October 2015; pp. 234–241.
54. Carreira, J.; Zisserman, A. Quo vadis action recognition? A new model and the kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4724–4733.
55. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]

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.