A lightweight deep learning approach to mouth segmentation in color images

Kittisak Chotikkakamthorn and Panrasee Ritthipravat
Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand
Worapan Kusakunniran
Faculty of Information and Communication Technology, Mahidol University, Nakhon Pathom, Thailand, and
Pimchanok Tuakta and Paitoon Benjapornlert
Department of Rehabilitation Medicine, Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand

Received 16 August 2022
Revised 28 October 2022
Accepted 16 November 2022

Abstract

Purpose – Mouth segmentation is one of the challenging tasks of development in lip reading applications due to illumination, low chromatic contrast and complex mouth appearance. Recently, deep learning methods effectively solved mouth segmentation problems with state-of-the-art performances. This study presents a modified Mobile DeepLabV3 based technique with a comprehensive evaluation based on mouth datasets.

Design/methodology/approach – This paper presents a novel approach to mouth segmentation by Mobile DeepLabV3 technique with integrating decode and auxiliary heads. Extensive data augmentation, online hard example mining (OHEM) and transfer learning have been applied. CelebAMask-HQ and the mouth dataset from 15 healthy subjects in the department of rehabilitation medicine, Ramathibodi hospital, are used in validation for mouth segmentation performance.

Findings – Extensive data augmentation, OHEM and transfer learning had been performed in this study. This technique achieved better performance on CelebAMask-HQ than existing segmentation techniques with a mean Jaccard similarity coefficient (JSC), mean classification accuracy and mean Dice similarity coefficient (DSC) of 0.8640, 93.34% and 0.9267, respectively. This technique also achieved better performance on the mouth dataset with a mean JSC, mean classification accuracy and mean DSC of 0.8834, 94.87% and 0.9367, respectively. The proposed technique achieved inference time usage per image of 48.12 ms.

Originality/value – The modified Mobile DeepLabV3 technique was developed with extensive data augmentation, OHEM and transfer learning. This technique gained better mouth segmentation performance than existing techniques. This makes it suitable for implementation in further lip-reading applications.

Keywords Mouth segmentation, Deep learning, MobileNetV2, Mobile DeepLabV3

Paper type Full length article

1. Introduction

Mouth segmentation is an important process in lip reading that can be applied in several applications, such as video conferencing, lip-synching, visual face recognition, speech

Funding: This research project is supported by Mahidol University (Basic Research Fund: fiscal year 2021) (FRB640032) (Contract No BRF1-011/2564).

Conflict of interest: The authors declare that there is no conflict of interest in this paper.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-08-2022-0225

recognition and medical disease detection [1–4]. The accuracy of each application depends on segmentation performances. However, mouth segmentation is challenging in an unconstrained environment due to luminance variation, low chromatic contrast, complex mouth appearance, fitness, occlusion, reflection and cosmetic agents on the lip [1–3, 5 and 6]. Currently, various techniques are used for separating the lip region from the background, such as contour-based and region-based approaches.

However, several challenges still exist, such as overlapping between the lip and non-lip color, and no obvious color gradient between lip and skin [3, 5 and 6].

After the introduction of AlexNet [7], the first end-to-end multi-resolution deep learning-based semantic segmentation technique is fully convolutional network (FCN) [8]. It achieves higher accuracy than conventional techniques. Later, mouth segmentation techniques have been continuously developed. Newer techniques can segment without color space transformation, manual feature extraction or even sliding window in pixel-wise prediction.

In this paper, we proposed a solution of the automatic deep learning-based mouth segmentation method evaluated on the publicly available dataset as CelebAMask-HQ [9], and the mouth dataset collected from 15 healthy people, annotated by four personnel in rehabilitation medicine, Ramathibodi hospital, and verified by two rehabilitation doctors. We applied transfer learning from COCO-Stuff [10] to CelebAMask-HQ dataset and applied from CelebAMask-HQ to the mouth dataset.

The key contribution of this paper is that we validated the performance of the Mobile DeepLabV3-based technique on mouth segmentation on the publicly available dataset and the self-collected mouth dataset from 15 healthy people. We integrated decode and auxiliary heads on Mobile DeepLabV3 to enhance supervision during training. This study applied extensive data augmentation and online hard example mining (OHEM) to relieve class imbalance. Our proposed model achieves better performance than the standard segmentation techniques. The second contribution is the application of transfer learning, taking the pretrained model on COCO-Stuff to CelebAMask-HQ, and the model trained on the CelebAMask-HQ to the mouth dataset, using a lesser amount of data for re-training. Moreover, our proposed solution does not require preprocessing and postprocessing. Thus, this can be easily integrated into mouth segmentation-related applications.

The rest of this paper is organized as follows: Section 2 describes the related work. Section 3 describes the materials and methods. Section 4 provides the experimental results. Section 5 discusses the results. Section 6 draws the conclusion.

2. Related work

Mouth segmentation techniques have been actively researched for solving an unconstrained condition on various illumination, mouth shape, reflection, and cosmetic agents on the lip [1–3, 5 and 6]. These can be separated into three categories: contour-based approach, region-based method and deep learning-based method.

First, the contour-based technique separates the lip and the background by a gradient between the lip and the non-lip pixels. Ozgur *et al.* [11] proposed PCA (Principal Component Analysis) template matching and K-means algorithm for lip corner detection. The likelihood of segmented lip pixels is estimated by Gaussian mixture model from the detected lip corner. Malek *et al.* [12] applied an active contour and parametric model to get lip contour. Then, a level set method finds the key points to position the result of the parametric model to fit lip deformity. Lu and Liu [2] proposed a localized active contour model from an illumination equalized RGB image, and the combination of the U component on the CIE-1975 CIELUV image, and C2 and C3 components from discrete Harley transformed image. This study applied the initial rhombus contour to the closed mouth and the combined semi-ellipses to the open mouth. Malek and Messaoud [13] proposed two methods. First, the authors proposed lip landmark detection by the geodesic active contour and a distance level set evolution model with a combination of Gaussian,

median and average filters [13]. Next, a parametric model based on the cubic curves estimates a lip deformity from a lip landmark [13].

Second, the region-based approach applies clustering or thresholding techniques to separate between a lip and a background. Sandhya *et al.* [14] applied Otsu's thresholding and K-means clustering from the grayscale lip-printed image. The separation of K-means clusters is based on Euclidean distance. Wang *et al.* [6] proposed multi-class and shape-guided fuzzy C-means (MS-FCM) from CIE-1975 CIELAB and CIELUV. The pixel vector from the selected channels L*, a*, b*, u* and v* was separated between the lip and the complex backgrounds like skin, beards and mustaches. Gritzman *et al.* [3] applied shape-based adaptive thresholding (SAT) through two processes. First, this study used linear discriminative analysis with support vector regression to output segmentation error. Next, this study adjusted the color-based threshold value to estimate the best value to reduce the segmentation error until it was acceptable.

The third approach is the deep learning-based technique. Ju *et al.* [5] proposed lip segmentation network (LSN) which combined features from two architectures. First, FCN-based architecture maps RGB to a binary image. Second, the proposed CNN architecture based on average pooling with a 1×1 convolution kernel is employed to reduce the bad annotation influence. Guan *et al.* [15] proposed lip segmentation fuzzy CNN (LSFCNN), the U-net-like architecture with fuzzy learning modules. Zhang and Zhao [16] proposed a U-net-based local feature extractor to extract visual information from lip images with complex environmental changes and different facial attributes. They also proposed a graph-based adjacent feature extractor to effectively capture features of lips between adjacent frames. Guan *et al.* [17] proposed LSDNet, the combination between complex teacher and student networks. This combines three loss functions: cross-entropy, distillery and remedy losses. LSDNet increases segmentation performance, inference speed and segmentation ability in hard samples.

Nowadays, little research applies end-to-end CNN with an auxiliary head, extensive data augmentation, OHEM and transfer learning to solve a mouth segmentation problem in an unconstrained condition. Moreover, no research studies on lip and teeth segmentation performance, and computational complexity. Then, this paper validates Mobile DeepLabV3-based techniques on lip and teeth segmentation performance and validates computational complexity by providing model parameters, model size and time usage per image.

3. Material and methods

3.1 Dataset

The first experiment was applied to CelebAMask-HQ [9], a large-scale publicly available high-resolution face dataset with fine-masked labels of 19 facial component categories such as eye, nose and mouth regions. CelebAMask-HQ has high-quality control from several rounds of verification and refinement of each annotated mask to reduce noise. The dataset contains 30,000 face images of 512×512 resolution.

The next experimental study was applied to the collected videos from 15 healthy people working in the department of rehabilitation medicine at Ramathibodi hospital. This experiment was approved by the institutional review board of Ramathibodi hospital, Mahidol University (certificate of approval (COA) number. MURA2021/73). The inclusion criteria were as follows:

(1) The subject requires to be a Thai.

(2) The subject should be between 18 and 80 years old.

(3) The subject works in the Faculty of Engineering at Mahidol University or the department of Rehabilitation Medicine at Ramathibodi hospital.

(4) The subject does not have a neck movement disorder or a history of cervical surgeries or trauma.

The exclusion criteria consist of a relationship with the research team, and unavailability during testing. Consent was obtained from all subjects for participating in the experiment.

The videos were acquired from the smartphone camera and the Razer webcam in an unconstrained environment in the department of rehabilitation medicine at Ramathibodi hospital. We extracted each video frame and saved it as a picture. The extracted frame was precisely annotated with Universal Data Tool (v.0.14.17) by four personnel working in the same department under the supervision of two rehabilitation doctors. Precise annotation with high-quality control and supervision reduces noise which affects training performance [5]. This dataset possesses 15,495 images.

3.2 CNN architecture

The model architecture used in this study is based on Mobile DeepLabV3 [18, 19]. It consists of three parts, i.e., the backbone, the auxiliary head and the decode head (Figure 1). First, the backbone architecture is derived from MobileNetV2 [19].

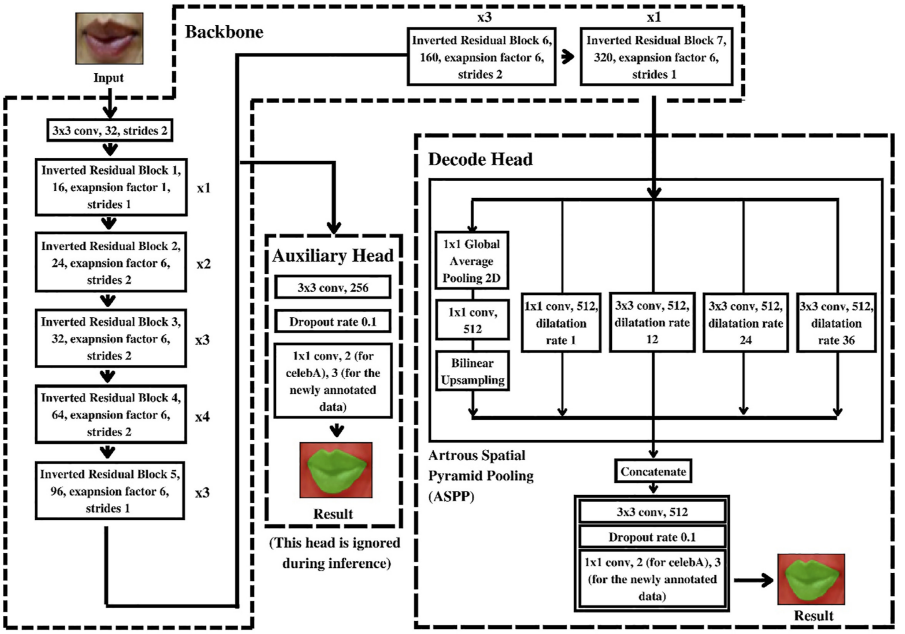
Second, an auxiliary head [20, 21] processes the output of the 5th inverted residual block for training optimization assistance. The reason is a vanishing gradient problem on the deeper network decreasing the gradient to near zero, which prevents fine-tuning parameters. Placing an auxiliary head on shallower layers increases a backpropagation signal, and additional regularization. Thus, an auxiliary head increases classifier performance from the insight of the InceptionNet [21]. This head consists of a 3×3 convolutional layer with an output channel of 256, a dropout rate of 0.1, and a 1×1 convolutional layer. This head outputs two classes for CelebAMask-HQ and three classes for the mouth dataset. This head is abandoned during inference.

Third, the decode head is the main architecture head that processes the output of the 7th inverted residual block to output the same classes of an auxiliary head. This head consists of four steps, i.e., Atrous spatial pyramid pooling (ASPP) [18, 22], a 3×3 convolution layer with an output channel of 512, a dropout rate of 0.1, and a 1×1 convolutional layer. ASPP is a powerful tool to capture semantic information on various scales from the computed feature maps by the model's receptive field enlargement. It consists of five parallel paths. The four parallel paths are Atrous convolutions [18, 22] containing three 3×3 convolution layers with different dilatation rates of 12, 24 and 36, and one 1×1 convolution layer with a dilatation rate of 1. The last path is the image-level feature extraction which has three processes: a 1×1 2D global average pooling, a 1×1 convolution layer with an output channel of 512, and a resize layer with bilinear interpolation to the same image resolution before passing through the image-level feature extraction. The output of five parallel paths is concatenated before passing through a 3×3 convolutional layer.

3.3 Methods

The first experiment was the assessment of CelebAMask-HQ [9] prepared by four processes. First, the lip area was cropped by taking the masked area between the upper and lower lips to take the coordinates of extreme points. Second, the dataset was reannotated by labeling the upper and lower lip-masked areas as the lip, and the others as the background. This dataset contains 29,928 background areas and 29,505 lip areas. Third, the reannotated images were resized to the resolution of 640×480 pixels. Last, the dataset was separated from 30,000 images into 20,950 training, 5,987 validation, and 2,991 testing images. A total of 72 remaining images were excluded due to errors during finding extreme points of the lip-masked areas which are not available.

This experiment applied Mobile DeepLabV3 [18, 19] pretrained by the COCO-Stuff dataset [10] to train with the training subset and validate with the validation subset. This network was trained with Adam optimizer for 140 epochs. The learning rate and weight decay were



Source(s): Refs. (18–22)

Figure 1.
The mobile DeepLabV3 segmentation technique

0.001 and 0.0001, respectively. The initial random seed was set to 0. We applied OHEM [23, 24] for the segmented pixel area with a confidence value of less than 0.7. OHEM filters the difficult segmentation pixels with a low confidence value for backpropagation. The neglected class during training provides a high loss enough until reaching the probability of being sampled. Thus, OHEM mitigates a large imbalance between the annotated objects and the background on the mouth dataset.

The loss function in the main and auxiliary heads is the combination of two components: cross-entropy (L_{CE}) and dice losses (L_{Dice}).

The cross-entropy loss (L_{CE}) is the sum of cross-entropy losses in every class between the ground truth (y_i) and prediction calculated by the softmax function of the normalized exponential function of the prediction value in the current class (p_i). The numerator is the sum of the exponential function of prediction values in each class (z_c). The total number of classes is represented as C. The cross-entropy loss is shown in equation (1).

$$L_{CE} = -\sum_{i=1}^{C} y_i log\left(\frac{e^{p_i}}{\sum_{c=1}^{C} e^{z_c}}\right). \quad (1)$$

The dice loss (L_{Dice}) is the average of the dice coefficient in every class. In each class, the sum of correctly predicted boundary pixels is the numerator, and the sum of the total boundary between the prediction and the ground truth is the denominator. p_i represents the pixel values of the prediction, and g_i represents the pixel values of the ground truth. N_c represents the number of pixels in each class. The total number of classes is represented as C. The dice loss is shown in equation (2).

$$L_{Dice} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\sum_{i}^{N_c} p_i g_i}{\sum_{i}^{N_c} p_i^2 + \sum_{i}^{N_c} g_i^2} \quad (2)$$

The final loss (L_{total}) for the main and auxiliary heads can be calculated as shown in equation (3).

$$L_{total} = L_{CE} + 3L_{Dice} \quad (3)$$

After the training, the model was tested with the testing dataset, and compared to the baselines (Part A of LSN [5], LSFCNN [15], LSDNet [17], U-Net [16, 25], FCN [8], PSPNet [20], Residual U-Net++ [26] and DeepLabV3 [18]) for segmentation accuracy. For LSN [5], only Part A was selected in this study because the author provided insufficient details on the structure of part B.

The next experiment was the assessment of the dataset collected from healthy people, containing 15,495 images, and preprocessed in three steps. Firstly, this dataset was annotated by the personnel working in the department of rehabilitation medicine to create the masked image in three classes: the lip, the teeth and the background areas. Then, this dataset contains 15,495 background areas, 15,487 lip areas and 4,894 teeth areas. Secondly, the annotated images were resized to 640 × 480 pixels. Lastly, the dataset was separated into 10,851 training, 3,097 validation and 1,547 testing images. The pretrained model from the previous experiment was applied to train with the training subset with the same training parameters as in the previous experiment and validate with the validation subset After training, the testing subset was used in the evaluation and compared to the baselines for segmentation accuracy except for LSFCNN and LSDNet. The main reason for the exclusion is that the model architecture and loss function were specially designed for lip segmentation, which was not flexible for including teeth.

Data augmentation [27] is applied on training sets of both datasets, used for all techniques to improve the sufficiency and diversity of training data by synthetic dataset generation. The model with data augmentation copes better with the variety of colors, illumination and geometric transformation. Data augmentation consists of three steps: random crop, random flip and photometric distortion which applied random brightness, random contrast, BGR-to-HSV conversion, random saturation, random hue, HSV-to-BGR conversion and random contrast.

The third experiment was an ablation study. The same method in second experiment is applied for the proposed model without ASPP, an auxiliary head, transfer learning and OHEM. The result was compared to the proposed model.

Three experiments were applied using PyTorch (v.1.11.0), mmCV (v.1.5.2) and mmSegmentation (v.0.22.1) deep learning libraries in Python (v3.8.10). A CUDA-enabled GPU (NVIDIA Geforce RTX 3060) with 12GB RAM was applied for training and testing processes.

The performance evaluation metrics in the three experiments' validation and testing phase were the mean Jaccard similarity coefficient (mean JSC), the mean classification accuracy and the mean Dice similarity coefficient (mean DSC).

The fourth experiment was the performance evaluation. We performed via Intel Core i7-4770 with a clock speed 4.50GHz and NVIDIA RTX 3060 to output the number of model parameters, the model size in MB and the inference time usage per image in milliseconds (ms).

4. Results

Figures 2 and 3 show the training and validation graphs of CelebAMask-HQ and the collected dataset from healthy people. This illustrates two learning graphs including training cross-entropy and dice losses, which shows early convergence on all training graphs since we

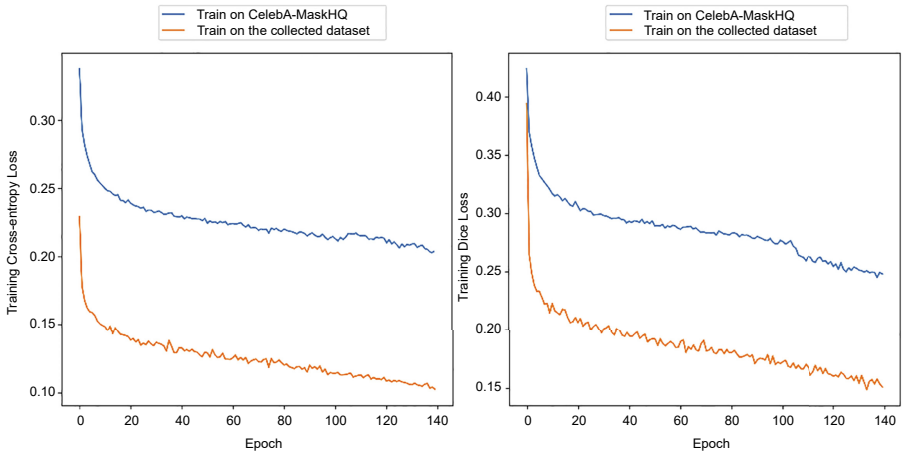


Figure 2.
The training graph on the training subset of CelebAMask-HQ and the collected dataset from 15 healthy people

applied transfer learning from COCO-Stuff, the large dataset, to CelebAMask-HQ, and from CelebAMask-HQ to the same domain on the mouth dataset. For validation on CelebAMask-HQ, the mean JSC mean classification accuracy and DSC achieved up to 0.8698, 93.66% and 0.9300, respectively. For validation on the mouth dataset, the mean JSC, mean classification accuracy and DSC achieved up to 0.8382, 93.39% and 0.9067, respectively.

The first experiment result on the testing subset of CelebAMask-HQ is shown in Table 1. Mobile DeepLabV3 demonstrated promising results, achieving mean JSC, mean classification accuracy, and mean DSC of 0.8640, 93.34% and 0.9267, respectively. The results demonstrated statistically significant improvement compared to the baselines ($p < 0.05$). An example of the ground truth images, labels and segmentation results is shown in Figure 4.

The second experiment result on the testing subset of the collected dataset is shown in Table 2. Mobile DeepLabV3 demonstrated promising results, achieving the mean JSC, classification accuracy and mean DSC of 0.8834, 94.87% and 0.9367, respectively. This technique demonstrated statistically significant improvement to the baselines ($p < 0.05$) except for DeepLabV3 on Mean JSC and DSC and residual U-Net++ on DSC. An example of the ground truth images, labels and segmentation results is shown in Figure 5.

The third experiment result shown in Table 4 is an ablation study of the testing subset of the created dataset. Mobile DeepLabV3 with ASPP, an auxiliary head, transfer learning approach and OHEM. Statistical analysis is applied to evaluate the significant difference between each study compared to the proposed model ($p < 0.05$).

Fourth, the performance evaluation result on mouth segmentation performance is shown in Table 3. Mobile DeepLabV3 has a lower number of parameters and a smaller model size than the baselines except for Part A of LSN [5], LSFCNN [15], LSDNet [17] and residual U-Net++ [26]. This model achieves faster inference time usage per image than the baselines except for Part A of LSN, LSFCNN and LSDNet.

5. Discussion

From our experimental results, Mobile DeepLabV3 outperforms Part A of LSN, LSFCNN, LSDNet, DeepLabV3, FCN, U-Net, PSPNet and residual U-Net++ with higher classification accuracy, higher mean DSC and mean JSC. Moreover, Mobile DeepLabV3 achieves the fastest segmentation speed except for Part A of LSN, LSFCNN and LSDNet.

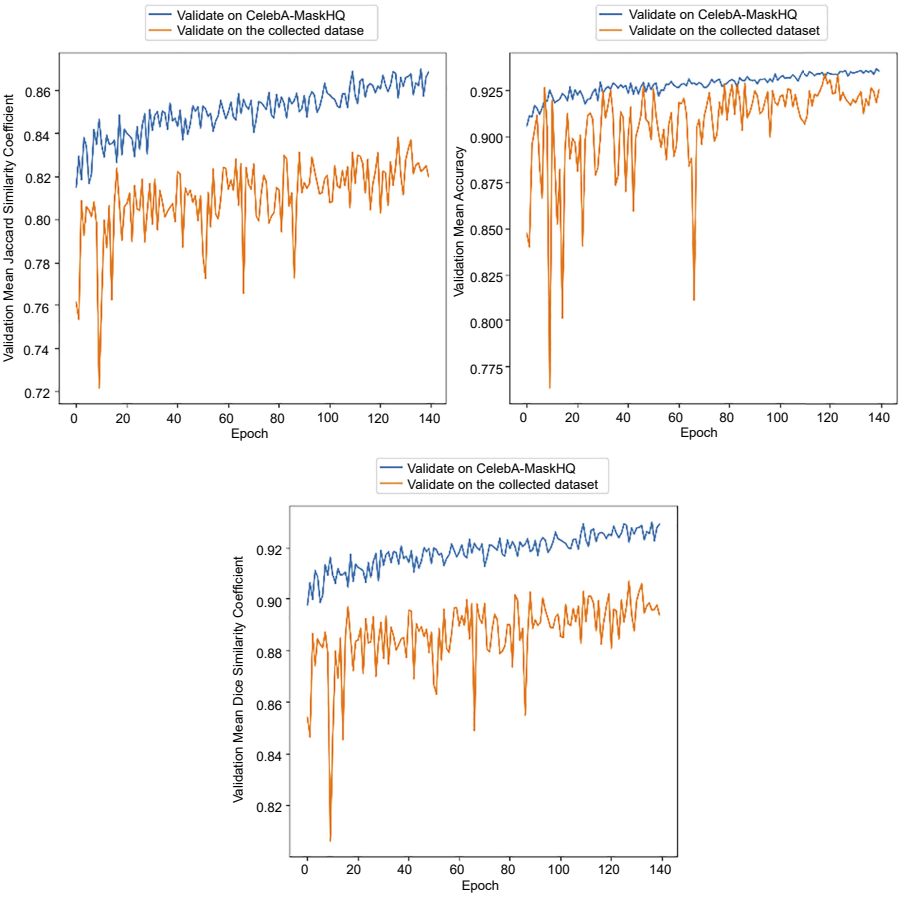


Figure 3.
The validation graph on the validation subset of CelebAMask-HQ and the collected dataset from 15 healthy people

Techniques	Mean classification accuracy (%)	Mean Jaccard similarity coefficient (mean JSC)	Mean Dice similarity coefficient (mean DSC)
Part A LSN [5]	81.95	0.6717	0.8022
LSFCNN [15]	82.93	0.7346	0.8431
U-Net [16, 25]	88.57	0.7825	0.8771
LSDNet [17]	92.51	0.8575	0.9250
FCN [8]	92.89	0.8543	0.9211
PSPNet [20]	92.50	0.8349	0.9097
Residual U-Net++ [26]	92.87	0.8535	0.9206
DeepLabV3 [18]	93.21	0.8612	0.9097
Mobile DeepLabV3 [18, 19]	93.34	0.8640	0.9267

Table 1.
The segmentation result of the testing subset of the CelebAMask-HQ dataset

Superior performance arises from five factors: ASPP [18, 22], an auxiliary head [20, 21], MobileNetV2 [19], OHEM [23, 24] and transfer learning [28].

First, ASPP [18, 22] enables capturing semantic information on various scales by the model's receptive field enlargement in the different dilatation rates on Atrous convolution [16, 20].

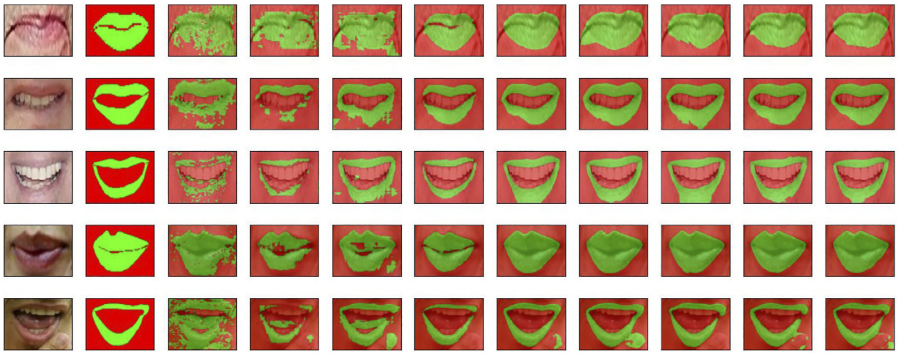
Second, an auxiliary head [20, 21] in the intermediate layer assists the training process by back propagation through the shallow layers. This prevents the gradient vanishing problem [20, 21].

The third factor is MobileNetV2 [19] including an inverted residual block that has a linear bottleneck. The bottleneck transfers the necessary information between residual blocks to decrease information and performance loss from the non-linearity transformation property from ReLU6.

Fourth, OHEM [23, 24], filters the difficult segmentation pixels with a low confidence value for backpropagation. The neglected class during training provides a high loss enough until reaching the probability of being sampled.

Fifth, the network-based transfer learning approach [28] applies the reusability and transferability properties of a trained deep-learning model. This mitigates a large amount of dataset requirement on a limited amount of data in the mouth dataset for training.

From Table 4, ASPP, an auxiliary head, OHEM and transfer learning lead to increase mouth segmentation performance compared to the model without these components. Moreover, from Tables 1–3, MobileNetV2 positively reinforces mouth segmentation accuracy with a reduction of computational complexity and memory requirement.



Note(s): The segmentation results show the green and the red areas which represent the lip and the background

Figure 4.
Ground truth images and labels, and segmentation results on the testing subset of the CelebAMask-HQ dataset are displayed in 11 columns: ground truth images, ground truth segmentation labels, segmentation results from Part A of LSN [5], LSFCNN [15], U-Net [16, 25], LSDNet [17], FCN [8], PSPNet [20], Residual U-Net++ [26], DeepLabV3 [18] and Mobile DeepLabV3 [18, 19]

Techniques	Mean classification accuracy (%)	Mean Jaccard similarity coefficient (mean JSC)	Mean Dice similarity coefficient (mean DSC)
Part A of LSN [5]	66.80	0.5354	0.6628
U-Net [16, 25]	80.47	0.7497	0.8470
FCN [8]	94.11	0.8764	0.9324
PSPNet [20]	93.18	0.8777	0.9334
Residual U-Net++ [26]	92.30	0.8612	0.9274
DeepLabV3 [18]	94.07	0.8841	0.9371
Mobile DeepLabV3 [18, 19]	94.87	0.8834	0.9367

Table 2.
The segmentation result of the testing subset of the collected dataset from healthy people

Figures 4 and 5, show satisfactory qualitative results of Mobile DeepLabV3 [16, 17]. However, Mobile DeepLabV3 still misclassified tongue, oral mucosa, skin and nail as the lip area due to color similarity to the lip, no obvious RGB color difference, low chromatic contrast, occlusion,

Figure 5.
Ground truth images and labels, and segmentation results on the testing subset of the collected dataset are displayed in 9 columns: ground truth images, ground truth segmentation labels, and segmentation results from Part A of LSN [5], U-Net [16, 25], FCN [8], PSPNet [20], Residual U-Net++ [26], DeepLabV3 [18] and Mobile DeepLabV3 [18, 19]

Note(s): The segmentation results show the green, red, and blue areas which represent the lip, the background, and the teeth, respectively

Table 3.
Performances of mouth segmentation regarding a number of parameters, a model size and inference time usage per image

Techniques	Parameters	Model size (MB)	Inference time per image (ms)
Part A of LSN [5]	44,247	0.55	47.17
LSFCNN [15]	10,332,236	39.50	45.40
LSDNet [17]	2,906,400	12.00	25.70
FCN [8]	49,485,414	568.30	78.00
U-Net [16, 25]	29,060,806	332.70	299.37
PSPNet [20]	48,963,174	560.80	74.74
Residual U-Net++ [26]	17,618,663	201.00	281.50
DeepLabV3 [18]	68,100,710	799.80	91.82
Mobile DeepLabV3 [18, 19]	18,589,702	213.10	48.12

Table 4.
The segmentation result on ablation studies of the testing subset of the collected dataset

ASPP	An auxiliary head	Transfer learning	OHEM	Mean classification accuracy (%) Value	*p*-value	Mean Jaccard similarity coefficient (mean JSC) Value	*p*-value	Mean Dice similarity coefficient (mean DSC) Value	*p*-value
✓	✓	✓	✓	94.87	–	0.8834	–	0.9367	–
✗	✓	✓	✓	93.88	0.00	0.8826	0.00	0.9362	0.00
✓	✗	✓	✓	90.64	0.00	0.8565	0.00	0.9195	0.00
✓	✓	✗	✓	94.31	0.00	0.8827	0.00	0.9364	0.00
✓	✓	✓	✗	93.80	0.00	0.8767	0.00	0.9327	0.00

Note(s): The first row provides the segmentation result from the proposed method

reflection and high variations of illumination and lip color [1–3, 5, 6]. These require further studies for dataset preparation and mouth segmentation model modification to improve segmentation performance.

Compared with the conventional techniques, this technique does not require preprocessing, and lip contour initialization and finding. This benefits from automatic feature extraction found in deep learning. It does not require additional conventional modules like fuzzy units [13] which increases computational complexity.

Compared to the baselines, Mobile DeeplabV3 [16, 17] is better. All baselines do not have MobileNetV2 as the backbone with an auxiliary head for supervision, and OHEM. Almost all baselines do not have ASPP and transfer learning except for DeepLabV3 [16] and LSDNet, respectively. The lack of segmentation performance improvement factors leads to deteriorating segmentation accuracy. Moreover, Part A of LSN [4], LSFCNN [13], and U-net-based techniques [14, 23, 24] performed worst. They achieved the lowest segmentation accuracy compared to the other baselines, and Mobile DeepLabV3 [16, 17]. They misclassified the inside and outside mouth areas as the lip and teeth.

6. Conclusion

In this paper, we proposed the mouth segmentation technique based on the Mobile DeepLabV3 technique to handle this problem by application of MobileNetV2 as the backbone architecture with the decode head based on ASPP and the auxiliary head, and with the extensive data augmentation, the application of OHEM to relieve a class imbalance problem and the transfer learning approaches from COCO-Stuff to CelebAMask-HQ, and from this dataset to the mouth dataset. Among the baseline techniques, the proposed method has been verified to be more accurate and faster in inference speed than others for the mouth segmentation problem. This technique is suitable for implementation in further lip-reading applications, visual face recognition, speech identification, video conference and medical disease detection.

References

1. Chowdhury DP, Kumari R, Bakshi S, Sahoo MN, Das A. Lip as biometric and beyond: a survey. Multimed Tools Appl. 2022; 81(3): 3831-65. doi: 10.1007/s11042-021-11613-5.
2. Lu Y, Liu Q. Lip segmentation using automatic selected initial contours based on localized active contour model. EURASIP J Image Video Process. 2018; 2018(1): 7. doi: 10.1186/s13640-017-0243-9.
3. Gritzman AD, Postema M, Rubin DM, Aharonson V. Threshold-based outer lip segmentation using support vector regression. Signal Image Video Process. 2021; 15(6): 1197-202. doi: 10.1007/s11760-020-01849-3.
4. Anantharaman R, Velazquez M, Lee Y. Utilizing mask R-CNN for detection and segmentation of oral diseases. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018: 2197-204. doi: 10.1109/BIBM.2018.8621112.
5. Ju Z, Lin X, Li F, Wang S. Lip segmentation with muti-scale features based on fully convolution network. 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). IEEE; 2018: 365-70. doi: 10.1109/DSC.2018.00059.
6. Wang SL, Lau WH, Liew AWC, Leung SH. Robust lip region segmentation for lip images with complex background. Pattern Recognition. 2007; 40(12): 3481-91. doi: 10.1016/j.patcog.2007.03.016.
7. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017; 60(6): 84-90. doi: 10.1145/3065386.
8. Shelhamer E, Long J, Darrell T. Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Machine Intelligence. 2017; 39(4): 640-51. doi: 10.1109/TPAMI.2016.2572683.

9. Lee CH, Liu Z, Wu L, MaskGAN LP. Towards diverse and interactive facial image manipulation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020: 5548-57. doi: 10.1109/CVPR42600.2020.00559.
10. Caesar H, Uijlings J, Ferrari V. COCO-stuff: thing and stuff classes in context. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018: 1209-18. doi: 10.1109/CVPR.2018.00132.
11. Ozgur E, Yilmaz B, Karabalkan H, Erdogan H, Unel M. Lip segmentation using adaptive color space training - sabanci University Research Database [Internet]; 2008 [cited 2022 Aug 15]. Available from: https://research.sabanciuniv.edu/id/eprint/10287/
12. Malek M, Anouar BMM, Aicha B. Automatic Lip segmentation with level set method. 2019 International Conference on Control, Automation and Diagnosis, ICCAD 2019 - Proceedings; 2019. doi: 10.1109/ICCAD46983.2019.9037912.
13. Miled M, Messaoud MAB. Lip segmentation with hybrid model. 2022 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP); 2022: 1-6. doi: 10.1109/ATSIP55956.2022.9805880.
14. Sandhya S, Fernandes R, Sapna S, Rodrigues AP. Segmentation of lip print images using clustering and thresholding techniques. In: Chiplunkar NN, Fukao T (Eds.), Advances in artificial intelligence and data engineering. Singapore: Springer Singapore; 2021: 1023-34. doi: 10.1007/978-981-15-3514-7_76.
15. Guan C, Wang S, Liew AWC. Lip image segmentation based on a fuzzy convolutional neural network. IEEE Trans Fuzzy Syst. 2019; 28(7): 1. doi: 10.1109/TFUZZ.2019.2957708.
16. Zhang C, Zhao H. Lip reading using local-adjacent feature extractor and multi-level feature fusion. J Phys Conf Ser. 2021; 1883(1): 012083. doi: 10.1088/1742-6596/1883/1/012083.
17. Guan C, Wang S, Liu G, Liew AWC. Lip image segmentation in mobile devices based on alternative knowledge distillation. 2019 IEEE International Conference on Image Processing (ICIP). IEEE; 2019: 1540-4. doi: 10.1109/ICIP.2019.8803087.
18. Chen LC, Papandreou G, Schroff F, Adam H. Rethinking atrous convolution for semantic image segmentation. arXiv. 2017. doi: 10.48550/arXiv.1706.05587.
19. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2018: 4510-20. doi: 10.1109/CVPR.2018.00474.
20. Zhao H, Shi J, Qi X, Wang X, Jia J. Pyramid scene parsing network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017: 6230-9. doi: 10.1109/CVPR.2017.660.
21. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2015: 1-9. doi: 10.1109/CVPR.2015.7298594.
22. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Machine Intelligence. 2018; 40(4): 834-48. doi: 10.1109/TPAMI.2017.2699184.
23. Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016: 761-9. doi: 10.1109/CVPR.2016.89.
24. Liu A, Wang Z. CV 3315 is all you need : semantic segmentation competition; 2022 [cited 2022 Aug 15]; Available from: https://arxiv.org/abs/2206.12571v2
25. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF (Eds.), Medical image computing and computer-assisted intervention – MICCAI 2015. Cham: Springer International Publishing; 2015: 234-41. doi: 10.1007/978-3-319-24574-4_28.

26. Jha D, Smedsrud PH, Riegler MA, Johansen D, Lange TD, Halvorsen P, Johansen HD. ResUNet++: an advanced architecture for medical image segmentation. 2019 IEEE International Symposium on Multimedia (ISM). IEEE; 2019: 225-2255. doi: 10.1109/ISM46123.2019.00049.

27. Yang S, Xiao W, Zhang M, Guo S, Zhao J, Shen F. Image data augmentation for deep learning: a survey; 2022 [cited 2022 Aug 15]; Available from: https://arxiv.org/abs/2204.08610v1

28. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. A survey on deep transfer learning. In: Kůrková V, Manolopoulos Y, Hammer B, Iliadis L, Maglogiannis I (Eds.), Artificial neural networks and machine learning – ICANN 2018. Cham: Springer International Publishing; 2018: 270-9. doi: 10.1007/978-3-030-01424-7_27.

Corresponding author
Panrasee Ritthipravat can be contacted at: panrasee.rit@mahidol.ac.th