Lip Segmentation with Muti-Scale Features Based on Fully Convolution Network

Zhen Ju
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
jixiandetian@sjtu.edu.cn

Xiang Lin
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China

Fangqi Li
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
solour_lfq@sjtu.edu.cn

Shilin Wang §
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, China
wsl@sjtu.edu.cn

***Abstract*—Recently, lip image segmentation has attracted much attention because the lip image segmentation result provides much visual information for recognition of visual speech. In this paper, a new approach for lip segmentation based on deep neural network is presented. The deep neural network is firstly introduced to lip image segmentation task while the current methods using the hand-crafted feature. In this method, muti-scaled information is integrated to do lip segmentation by using various size convolution kernels. Two neural networks is combined in this method which reduce the bad effect of dirty annotation. Also a new loss function is employed, which take full advantage of the knowledge between successive frames, so that it is more robust to the dirty annotation and unbalanced data. Our method can achieve good performance using the low-quality annotation. Experimental results show that this method achieves the state-of-art performance.**

***Index Terms*—lip segmentation, neural network**

I. Introduction

The visual information of the lips can enhance the performance of speech recognition and identity authentication based on the lips [1]–[7]. The visual information of lips contains lip shape, lip color, lip segmentation and so on. Accurate lip segmentation is of cardinal significance for application based on the lips. However, the lack of precise annotation and the low color contrast between lip and non-lip region makes the problem difficult.

Recently, many techniques have been proposed in the literature. There are three major kinds of segmentation approach. The first kind is to perform the lip segmentation by active contour algorithm or active shape model. Delmas et al. [8] appied active contour algorithm to lip segmentation task, but that method leaked robustness to the initial position, it would produce unsatisfactory segmentation results while the initial position is far from lips' edge. Moreover it is hard to find appropriate initial value in most cases. Also active shape model is widely used [9], but this method is sensitive to the environment. It produces inaccurate results when the boundary between lip and non-lip regions is not obvious, especially the lip color is close to the face or tongue color.

Clustering method is another kind of the lip segmentation techniques. Various methods have been proposed based on classical fuzzy c-means (FCM) algorithm [7], [10]. Also there are many improved methods in the past years in order to enhance the performance of FCM [11]–[16]. However those methods also perform poorly when processing the image of week color contract. Leung et al. [17] has improved this method by adding spatial constraint, this phenomenon has been relieved but also exists.

The last kind is to perform the segmentation directly from the color space [18], [19]. Preprocessing is necessary in this kind of algorithm in order to amplify the difference between lip and non-lip region, the preprocessing includes color transformation, color filter and so on. This kind of method is very fast but it is also sensitive to the environment. While processing the image with weak color contrast, this method would have poor performance because the method can not distinguish the boundary between lip and non-lip regions accurately. There exist improved methods [20], [21], which have been proposed based on the markov random field technique. The improved methods can enhance the robustness and reduce the segmentation error. However, those methods generally make mistakes when processing the pixel inside the mouth region.

Recent years, fully convolutional network (FCN) [22] is proposed for semantic segmentation task, and several improved methods based on FCN are proposed. However, FCN is sensitive to the noise especially with several given patterns and it will not work well using the FCN directly with the data annotation of low qualify. But it is expensive to label all the train images precisely. Therefore it is very meaningful to explore a new method which is robust to the annotation of low quality. A lip segmentation network (LSN) is designed

§ Shilin Wang is corresponding author.

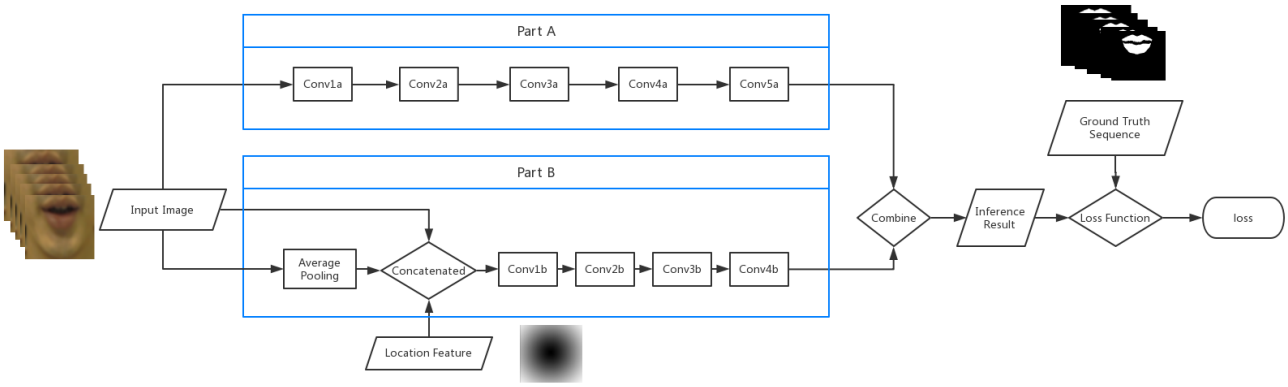

Fig. 1. The proposed network structure

to solve this problem. The proposed network combines two sub-network, one adopts a large size kernel so this sub-network has a big perceptual field, while the other mainly uses 1×1 kernel to fully use the local information and avoid the over-fitting caused by the noise pattern. There is relationship between successive frame of a video, which can reduce the bad influence of dirty annotation. So the knowledge between successive frames is exploited by designing a new loss function.

The contributions of the proposed approach are summarized in the following:

- The deep neural network is firstly exploited to do the lip segmentation task.
- A new light network is proposed for lip segmentation. This network combine muti-scale information and it can perform well even with annotation of low quality.
- The approach presents a new loss function that can use the information between successive frame of video. In this way, we can reduce the bad effect of the inaccurate annotation.

In Section II, details of the proposed method are presented. Section III presents the implementation of the lip segmentation. In Section IV, we show the experiment results. Section V draws the conclusion.

II. THE PROPOSED METHOD

Lip segmentation is a specialized semantic segmentation task. Recently, the deep convolution network technique has been introduced to the semantic segmentation, and those methods achieve the start-of-art performance. So we want to introduce this technique to lip segmentation task. However, lip segmentation is a relatively simple task and the existed network is too heavy for the lip segmentation task and sensitive to noise. Therefore a light and more robust network is necessary for the lip segmentation task.

In this paper, a light and more robust network named lip segmentation network (LSN) is proposed to process the lip segmentation. The training and testing data are continuous frames in videos. It is expensive to label all the pixel of images into lip and non-lip region accurately, so we obtain the annotation of data by unsupervised method and available tools. In this way, the pixel-level annotation of images can easily obtained, but the annotation is of low quality and the mistake is of several fixed patterns. The erroneous annotation usually occurs in the region of the tooth. The proposed approach reduces the bad influence of the low-quality annotation significantly.

A. Structure of Proposed Network

FCN is a powerful tool for segmentation task, however it is sensitive to the annotation and would not work well with erroneous annotation of several given patterns. So a new structure is proposed to solve this problem. Fig. 1 shows the structure of the proposed method. The whole network can be separated into two sub-network: PartA and PartB. The PartA adopts the idea of FCN , but is much lighter than FCN. This part will map the input color image to a binary image of the same size, the mapping function can be described by (1)

$$C = M_1(I) \tag{1}$$

where I is the input three-channel image, C is the output binary image of the same size, if the pixel value is zero, the pixel is classified as non-lip region. M_1 is the mapping function.

The patterns which the erroneous annotation has would be learned by PartA, so we design PartB to solve this problem. The PartA learns the erroneous patterns using the information of adjacent pixels. Therefore we mainly use 1×1 convolution kernel to restrict the relationship among the adjacent pixels. We can not judge exactly whether a pixel belongs to lip or non-lip region just using the color information, because the color of lip region is various in different environment. With the information of the adjacent pixel, it gets easier to judge. We should exploit several simple local information, so the average value of adjacent pixels is added by using the average pooling layer. According to the priori knowledge that the mouth is always in the center of the whole image, the

location information is added to improve the performance. The mapping function of PartB can be described by (2)

$$P = M_2(I, Loc) \tag{2}$$

where Loc is a single-channel image which contains the location information, P is the output matrix , its pixel value represents the probability that the pixel belongs to lip region. M_2 is the mapping function. The Loc can be computed by (3)

$$Loc_{i,j} = \max\{|\frac{size}{2} - i|, |\frac{size}{2} - j|\} \tag{3}$$

where $size$ means the width and height of the image.

Two parts will be integrated to predict the segmentation results. The combination function is described by (4)

$$S = Combine(C, P) \tag{4}$$

where $Combine$ is the combination function, detail description of (4) is shown as (5).

$$S_{i,j} = \begin{cases} 0 & \text{if} \quad C_{i,j} = 0 \\ 0 & \text{if} \quad C_{i,j} = 1 \; and \; P_{i,j} < thre \\ 1 & \text{if} \quad C_{i,j} = 1 \; and \; P_{i,j} \geq thre \end{cases} \tag{5}$$

where $thre$ is the threshold that whether the pixel is belonging to lip region.

B. The Improved Loss Function

The segmentation task is considered as pixel-level classification task, the definition of loss function for segmentation task is generally defined as (6)

$$LS(S, L) = \sum_{i,j} CE(L_{i,j}, S_{i,j}) \tag{6}$$

where S represents the inference result, L represents the ground truth, CE is a cross entropy loss function which is commonly used in classification task which can be described as (7).

$$CE(p_1, p_2) = -p_1 \log p_2 - (1 - p_1) \log (1 - p_2) \tag{7}$$

In consideration of the unbalance pixel number of lip and non-lip region, we improve the cross entropy loss function by introducing a weighting factor. In this way, we can reduce the bad influence of the unbalanced pixel number. The weight-balanced loss function is described by (8)

$$CE(p_1, p_2) = -p_1 \log p_2 * \alpha - (1 - p_1) \log (1 - p_2) \tag{8}$$

where α is the pixel number ratio of non-lip region and lip region.

By introducing the weight-balance cross entropy loss function, the definition for segmentation loss function would be described as (9)

$$LS_b(S, L) = \sum_{i,j} CE_b(L_{i,j}, S_{i,j}) \tag{9}$$

The data used in lip segmentation task is continuous frames in videos. Supposing that the video V consists of image sequence $\{Fr_1, Fr_1 \cdots, Fr_n\}$, $\{L_1, L_1 \cdots, L_n\}$ is the annotation images of the image sequence, $\{S_1, S_1 \cdots, S_n\}$ is the inference results of the image sequence. There is relationship between adjacent frames. The adjacent frames are almost same, and some erroneous annotation may not occur in all adjacent frames. Based on this observation, a new fuzzy loss function is proposed to reduce the bad influence of error annotation. The improved loss function is described by (10)

$$L_v(S_i, V) = \min\{LS_b(S_i, L_{i+j})\}, i \in \{-k, \cdots, k\} \tag{10}$$

where the k is the range to search the adjacent frames. If $i + j \leq 0$, $LS_b(S_i, L_{i+j})$ will be calculated as $LS_b(S_i, L_1)$. If $i + j > n$, $LS_b(S_i, L_{i+j})$ will be calculated as $LS_b(S_i, L_n)$.

The difference between inference results and the adjacent annotation is calculated, and the minimum one is selected as the difference between inference results and ground truth. In this way, the bad influence of erroneous annotation will be reduced. By employing the improved loss function LS_b, the proposed method is more robust to class imbalance and achieves better performance.

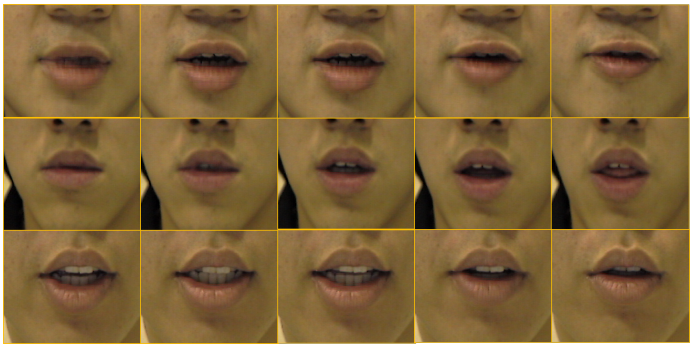

Fig. 2. Sample lip images in our dataset

III. IMPLEMENTATION

A. Dataset and Annotation

Lip images are captured by a video camera in RGB color format. The dataset that this paper used contains $36K$ images, the $36K$ images belong to 40 persons. The $36K$ images are separated into testing data and training data, testing data contain $9K$ images which belong to 10 persons, while the training data contain $27K$ images which belong to 30 persons. The original images contain the whole face, and we will preprocess the images. The images containing the whole face is cropped into images containing the lip area. Some sample lip images after preprocessing are illustrated in Fig. 2.

The outline of the lip region is labeled by 14-points. There is no annotation for the inner contour of the lips, so we adopt a unsupervised method to detect the inner contour. According to Matas et al. [22], an extremal region is a connected area of an image whose pixels have either higher or lower intensity than its outer boundary pixels, the inner contour boundary satisfies this condition. It is found that the maximally stable extremal regions algorithm (MSER) is able to detect almost all the inner

(a) (b)

Fig. 3. Sample of erroneous annotation.Figure (a) represents the origin lip images, figure (b) represents the annotations of the origin lip images.

contour. However, Fig. 3 shows the erroneous pattern that annotation has. As we can see, the tooth region usually be classified as lip region and there is generally breakage in the lip region.

B. Data Augmentation

It is found that the model is sensitive to the illumination. To make the model more robust to various illumination and background, each training image is randomly augmented by the following options:

- Adjust the contrast of input images by a random factor. In the HSV color space, the saturation S and V luminance components are changed, and the hue H is kept unchanged. The S and V components of each pixel are exponentially operated.
- Adjust the brightness of input images by a random factor.

Fig. 4. Sample of inference results, the first row represents the origin lip image, the second row represents the annotations, and the third row represents the inference results.

C. Detail of the Proposed Network

The network architecture of the proposed method is presented in Fig. 1, It is shown that the network contains two sub-networks, PartA and PartB.

The kernels of PartA are of the size 3×3, which are set empirically to achieve the best performance. There are five convolution layer in PartA, a stride of (1, 1, 1) is set for the convolution layer. Each convolution layer is followed by a relu layer. The output feature map for the five convolution layer is 16, 32, 64, 32, 2 respectively. There is no pooling layer in PartA.

There are four convolution layer in PartB, whose kernels is of the size 1×1. The average value of the adjacent pixel is implemented using a ave-pooling layer, whose stride is (1, 1, 1) and kernels size is 7×7. The output feature is concatenated to the first convolution layer, also a feature map contain the location information is concatenated to the first convolution layer.

Batch normalization layer is applied after each convolution layer to speed up the training process and achieve better performance. The training of PartA and PartB is independent. $thre$ in (5) is set 0.41 empirically to achieve the best performance compared with other value. Also k in (10) is set 2 empirically.

IV. EXPERIMENTAL AND DISCUSSION

A. Segmentation Result

In this experment, we use 3 images to demonstrate the segmentation results of our proposed method. The external and internal contour of the 3 images is annotated accurately. We compare the segmentation results of our method (LSN), color transformation algorithm (CT) [18], FCM, robust fuzzy c-means clustering algorithm (RFCM) [11], Lievin and Luthons method (LL) [20], and Zhang and Mercereaus method (ZM) [21], Leungs method (FCMS) [17] and FCN [22]. We also compare LSN with only Part A (LSN-A), our method without the improved loss (LSN-SL). Moreover, we experiment on a further 27 lip images to provide a more comprehensive evaluation for different methods.

The Segmentation Error is applied to evaluate the quality of segmentation results for lip images, which is defined as [24]

$$SE = P(O) \cdot P(B|O) + P(B) \cdot P(O|B) \qquad (11)$$

where $P(B|O)$ is the probability of classifying background as object, $P(O|B)$ is the probability of classifying object as background. $P(O)$ and $P(B)$ is the priori probabilities of the object and the background of image respectively.

a) Comparison of different methods: We compare the performance of LSN, CT, FCM, RFCM, LL, ZM and FCN. The results are shown in Table. I and Table. III. From the two tables, we can see that our method achieves the start-of-art performance.

TABLE I
PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS

Table	**Image 1**			**Image 2**			**Image 3**		
	$P(B\|O)$	$P(O\|B)$	SE**(%)**	$P(B\|O)$	$P(O\|B)$	SE**(%)**	$P(B\|O)$	$P(O\|B)$	SE**(%)**
CT	0.317	0.005	9.56	0.093	0.010	2.53	0.011	0.090	8.18
FCM	0.028	0.061	5.13	0.133	0.021	4.15	0.765	0.001	8.90
RFCM	0.109	0.031	5.39	0.818	0.012	16.23	2.038	0	23.59
FCMS	0.017	0.036	3.04	0.067	0.012	2.27	0.121	0	1.40
LL	0.007	0.104	7.60	0.247	0.004	4.94	1/054	0.004	11.36
ZM	0.048	0.045	4.58	0.023	0.039	3.61	0.135	0.035	4.56
FCN	0.018	0.034	2.94	0.042	0.008	1.4	0.094	0.003	1.08
LSN	0.011	0.021	**1.81**	0.023	0.004	**1.1**	0.046	0	**0.53**

TABLE II
PERFORMANCE COMPARISON BETWEEN STRUCTURE AND LOSS

Table	**Image 1**			**Image 2**			**Image 3**		
	$P(B\|O)$	$P(O\|B)$	SE**(%)**	$P(B\|O)$	$P(O\|B)$	SE**(%)**	$P(B\|O)$	$P(O\|B)$	SE**(%)**
LSN	0.011	0.021	**1.81**	0.023	0.004	**1.1**	0.046	0	**0.53**
LSN-A	0.019	0.031	2.75	0.042	0.010	1.41	0.073	0.004	0.85
LSN-SL	0.014	0.024	2.11	0.032	0.007	1.16	0.063	0.002	0.73

TABLE III
PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS

Method	***Average Segmentation Error (%)***
CT	4.38
FCM	4.13
RFCM	9.20
FCMS	2.72
LL	7.45
ZM	5.04
FCN	2.83
LSN	**1.89**

TABLE IV
PERFORMANCE COMPARISON BETWEEN STRUCTURE AND LOSS

Method	***Average Segmentation Error (%)***
LSN	**1.89**
LSN-A	2.98
LSN-SL	2.13

b) Effect of different structure and loss function: We compare the performance of LSN, LSN-A, LSN-SL to illuminate the effect of different network and loss function. The results are shown in Table. II and Table. IV. From the two tables, the average SE has been reduced by 0.24 percent after introducing the improved loss function. By integrating PartB, there is 1.09 percent reduction for average SE.

B. Robustness Analysis

a) Robustness to the Annotation: There are several erroneous annotation pattern which is caused by the MSER algorithm, we do several experiment to test whether the proposed method learn the erroneous patterns. The sample of segmentation results is shown in Fig. 4, the testing images are of erroneous annotation. From the experiment result, it is shown that the proposed method can resist against the erroneous annotation. As shown in Fig. 4, the proposed method performs well for the lip images with erroneous annotation.

b) Generalization of Model: To provide a generalization evaluation of the proposed method, we test our model using another dataset which is of different illumination and captured with different camera. 30 images of that dataset are labelled the external and internal contour. The average SE for the 30 images is 2.11%, which demonstrates that our model have good generalization.

V. CONCLUSIONS

In this paper, a new lip segmentation network is proposed. By the integration of muti-scale information, the proposed method can achieve good performance even if there exists erroneous annotation. Considering the knowledge between adjacent frames and class imbalance, the improved weight-balance loss function is proposed. By introducing the improved loss function, our approach reduces the bad influence of annotation of low quality and class imbalance. Moreover, the proposed method is robust to various illumination and background by data augmentation. From the experiment results, our approach achieves the start-of-art performance.

ACKNOWLEDGMENT

The work described in this paper was fully supported by National Natural Science Foundation of China (61771310), and program of Shanghai Technology Research Leader under grant 16XD1424400.

REFERENCES

[1] E. D. Petajan, "Automatic lipreading to enhance speech recognition, in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 1985, pp. 4047.

[2] N. P. Erber, "Interaction of audition and vision in the recognition of oral speech stimuli, J. Speech Hearing Res., vol. 12, pp. 423425, 1969.

[3] Y. Zhang, S. Levinson, and T. Huang, "Speaker independent audio-visual speech recognition, in Proc. IEEE Int. Conf. Multimedia and Expo, vol. 2, New York, July 2000, pp. 10731076.

[4] G. Rabi and S. Lu, “Visual speech recognition by recurrent neural networks, in Electrical and Computer Engineering, 1997. Engineering Innovation: Voyage of Discovery St. Johns, Nfld., Canada, May 1997, vol. 1, pp. 5558.

[5] J. Luettin, N. A. Thacker, and S. W. Beet, “Visual speech recognition using active shape models and hidden Markov models, in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 2, Atlanta, GA, May 1996, pp. 817820.

[6] C. Benoit, T. Mohamadi, and S. Kandel, “Effects of phonetic context audio-visualintelligibilityofFrench,J.SpeechHearingRes.,vol.37, pp. 11951203, 1994.

[7] J. C. Bezdek, Pattern Recognition With Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

[8] P . Delmas, P .Y . Coulon, V . Fristot, “Automatic Snakes For Robust Lip Boundaries Extraction, IEEE International Conference On Acoustic, Speech, and Signal Processing (ICCASP99), Phoenix, USA, 1999.

[9] J. Luettin, N.A. Tracker, S.W. Beet, “Active Shape Models for Visual Speech Feature Extraction, Electronic System Group Report N95/44, University of Sheffield, UK, 1995.

[10] J. C. Bezdek, “A convergence theorem for the fuzzy ISODATA clustering al- gorithms, IEEE Trans. Pattern Anal. Machine Intell., vol. 2, pp. 18, 1980.

[11] P.R.Kersten,R.Y.Lee,J.S.Verdi,R.M.Carvalho,andS.P.Yankovich, “Segmenting SAR images using fuzzy clustering, in Proc. 19th Int. Conf. North American, Fuzzy Information Processing Society, Atlanta, GA, July 2000, pp. 105108.

[12] P.R.Kersten, “Fuzzyorderstatisticsandtheirapplicationtofuzzyclus- tering, IEEE Trans. Fuzzy Syst., vol. 7, pp. 708712, Dec. 1999.

[13] T. A. Runkler and J. C. Bezdek, “Image segmentation using fuzzy clustering with fractal features, in Proc. 6th IEEE Int. Conf., vol. 3, Barcelona, Spain, July 1997, pp. 13931398.

[14] Y. A. Tolias and S. M. Panas, “On applying spatial constraints in fuzzy image clustering using a fuzzy rule-based system, IEEE Signal Processing Lett., vol. 5, pp. 245247, Oct. 1998.

[15] Y. T. Qian and R. C. Zhao, “Image segmentation based on combination of the global and local information, in Proc. Int. Conf. Image Processing, vol. 1, Santa Barbara, CA, Oct. 1997, pp. 204207.

[16] A. W. C. Liew, S. H. Leung, and W. H. Lau, “Fuzzy image clustering incorporating spatial continuity, Proc. Inst. Elect. Eng., vol. 147, no. pp. 185192, Apr. 2000.

[17] Leung S H, Wang S L, Lau W H. , “Lip image segmentation using fuzzy clustering incorporating an elliptic shape function”. IEEE transactions on image processing, 2004, 13(1): 51-62.

[18] N. Eveno, A. Caplier, and P. Y. Coulon, “New color transformation for lips segmentation, in Proc. IEEE 4th Workshop on Multimedia Signal Processing, Cannes, France, Oct. 2001, pp. 38.

[19] T. Wark, S. Sridharan, and V. Chandran, “An approach to statistical lip modeling for speaker identification via chromatic feature extraction, in Proc. 14th Int. Conf. Pattern Recognition, vol. 1, Brisbane, Australia, Aug. 1998, pp. 123125.

[20] M.LievinandF.Luthon “Unsupervisedlipsegmentationundernatural conditions, in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 6, Phoenix, AZ, Mar. 1999, pp. 30653068.

[21] X. Zhang and R. M. Mersereau, “Lip feature extraction toward an automatic speechreading system, in Proc. IEEE Int. Conf. Image Processing, vol. 3, Vancouver, BC, Canada, Sept. 2000, pp. 226229.

[22] Long, Jonathan, Evan Shelhamer, and Trevor Darrell, “Fully convolutional networks for semantic segmentation.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[23] J. Matas, O. Chum, M. Urban, and T. Pajdla, “Robust wide baseline stereo from maximally stable extremal regions, in Proc. Brit. Mach. Vis. Conf., vol. 1. 2002, pp. 384393.

[24] S. U. Lee, S. Y. Chung, and R. H. Park, “A comparative performance study of several global thresholding techniques for segmentation, Comput. Vis., Graph., Image Process., vol. 52, pp. 171190, 1990.