Accuracy and Speed Improvement in Lips Region Extraction

Takuya Akashi
Graduate School of Engineering
University of Tokushima
2-1 Minami-josanjima, Tokushima, Japan
akataku@is.tokushima-u.ac.jp

Minoru Fukumi
Faculty of Engineering
University of Tokushima
2-1 Minami-josanjima, Tokushima, Japan
fukumi@is.tokushima-u.ac.jp

Norio Akamatsu
Faculty of Engineering
University of Tokushima
2-1 Minami-josanjima, Tokushima, Japan
akamatsu@is.tokushima-u.ac.jp

Abstract

In this paper, a method for improvement of extraction accuracy and speed of lips region is described. In our previous study, we tried to extract lips region whose shape is varied by speech. Moreover, we want to apply this system in a mobile devices. Therefore, we think the condition is a nonrigid scene which is taken by a nonstationary camera. To solve these problem, we use template matching and the matching process is performed using a genetic algorithm. In addition, we think a characteristic shape and colour of the lips are effective for lips region extraction. Furthermore, to make the exploration of a genetic algorithm more efficient, we improve the fitness function and genetic operators. In our simulations, we obtain more efficient exploration, high accuracy and very short speed processing time. In this paper, these method of improvement and comparisons of results between our conventional method and improved method are described.

1. Introduction

Lately, the population that uses internet or e-mail by mobile devices, especially a cellular phone is increasing. However, the interface to input text is not convenient, because a cellular phone has only some key buttons for dialling. Therefore, we think that speech recognition is useful as an interface. In the field of speech recognition, there is a limit to recognize [4], in particular, a paragraph including a person name and a proper name. In the meantime, almost all cellular phones contains a small and high quality digital camera unit. Therefore, to improve the recognition rate, the system which we aim to achieve is translation of speech to text by using both speech and lips image data.

To make the process easy, the purpose of our study is to extract the lips region as a preprocessing of the system. However, this extraction process has some problems. The first is a nonrigid scene. Because the face move constantly and the camera is nonstationary and shake with the hand, the whole scene has considerable change. The second is the varying shape of lips by speech. The third is the processing time must be short. To solve these problems, we use template matching and the matching process is performed using a genetic algorithm with a characteristic shape and colour of the lips. In this paper, the method for improvement of our method [2] and comparisons of results between our conventional method and improved method are described.

2. Methodology

One of the earlier work is personal identification by using the feature of lips motion is described in [5]. This method calculates Mahalanobis distance from varying lips shapes by speech, but if the wrong feature points are detected, there is possibility of failure to recognize the lips shape. Also lipreading by using Eigen-template as the feature of the lips shape is reported in [8]. But, these works have a problem to being unadaptable for intensified geometric changes by a slope of a face or a camera and by a non-stable camera. That is the reason why one of the assumption is the camera and head of the subject are fixed.

On the other hand, to take the shape of the object into consideration, methods that use template matching have

been proposed [1, 3, 6, 10, 11]. In case the matching target has geometric changes that are parallel translation, scaling and rotation, the most basic approach is using multiple templates. But this approach has two problems. One, the more the number of templates used, the more template matching processing is required, which is time-consuming. Two, the fewer the number of the templates become, the less the extraction accuracy is, because few number of the templates means that the number of quantisation steps is large. For these problems, a template-matching method that constructs a parametric template space from a given set of template images is proposed [11]. But this method cannot apply to our study in terms of usefulness for our purpose, because a set of template images have to be prepared depending on the varying object shapes. To deal with the problems, a template matching using a genetic algorithm as a matching process is used in our proposed method. A large number of researches have been carried out into applications of the template matching using a genetic algorithm [3, 6, 10]. But, as far as we know, there is not a method that has invariance by only one template for shape deformations of lips that are an opened or closed mouth and showing or not showing any teeth, at the moment of speech.

Our conventional method [2] (SF-NPD: Static Fitness function with Normalised Pixel Difference) has invariance for an opened and closed mouth showing or not showing any teeth, and has high speed and high extraction accuracy using only one template by utilisation of the colour of lips and characteristics of shape deformations of lips during speech. Moreover, some distinctive inventions on a genetic algorithm make extraction processing relatively stable and high speed. In this paper, the proposed method (DF-PD: Dynamic Fitness function with Pixel Difference) which improves the SF-NPD method is described.

2.1. System Flow

As mentioned in 1, the lips region must be extracted notwithstanding heavy geometric changes and shape deformations in every frame of movie data. We, therefore, use template matching and the matching process is performed by a genetic algorithm. The system flow which we propose is as follows:

Step 1: Input a template and target image.

Step 2: Deform the template shape to a "square annulus".

Step 3: Generate a population of individuals of the first generation randomly.

Step 4: Measure fitness of individuals.

Step 5: Perform genetic operators.

Step 6: Output image which has extracted lips region.

In Step 1, image data is obtained from input images which are inputted on by one as test for our system. A x component (redness) of the Yxy colour space [7, 9] is used as this image data. The template shape is normal square in Step 1. By this normal square template can not support the varying lips shapes such as an opened or closed mouth and showing or not showing any teeth, at the moment of speech. Therefore, the template shape is deformed to a new shape called "square annulus" (refer to 3.1) in Step 2. From Step 3 to Step5, a genetic algorithm is performed. Step4 is described below in detail. Finally, a result of extraction is output on the target image in Step 6.

The procedure in Step 4 is as follows:

Step 4-1: Transform a template by using homogeneous coordinates.

Step 4-2: Control the template aspect ratio.

Step 4-3: Calculate a distance by the pixel difference between the template image and the target image.

Step 4-4: Penalize the distance according to the pixel difference.

Step 4-5:

SF-NPD: Calculate a fitness function and measure.

DF-PD: Calculate an objective function and a dynamic fitness function. and measure.

In Step 4-1, each individual parameters are obtained from informations based on the chromosome. These parameters are used for transforming the template by using homogeneous coordinates in Step 4-1. During the transformation, the aspect ratio of the template is controlled by Step 4-2. Then, in Step 4-3, a distance is calculated between the template image and the target image. And the distance is penalised according to the pixel difference in Step 4-4. This penalty is described in see 4.2. Then, in Step 4-5, in case of the SF-NPD method, fitness function is calculated using genetic operators. Then, the fitness value is measured (refer to 4.2). In the other case, the DF-PD method, at first an objective function is calculated using genetic operators. Next, fitness function is calculated and measured (refer to 4.2). The SF-NPD method has only a fitness function. This means that the objective function equals to the fitness function. This fitness value allows us to do a good exploration that the fitness approaches to 1. On the other hand, in the DF-PD method, an objective function and fitness function are distinguished.

3. Features of Varying Lips Shapes

To achieve invariance for shape deformations of lips, that is, an opened or closed mouth and showing or not showing any teeth, at the moment of speech, and to achieve high speed and high recognition accuracy by using only one template, we use some features of the varying shape of lips during speech. In this section, these methods are explained.

3.1. Shape of Template

In general, the typical template shape is a square. However, considering an application of the template matching to varying lips shapes, the square shape of template is unsuitable. This is because, at the moment of speech, the lips region has intense variations such as an opened or closed mouth and showing or not showing any teeth. In other words, the lips shape changes to other shapes constantly during speech. This is a serious problem in extraction of lips region by using only one template per user.

To solve this problem, we focus our attention on invariance under constantly varying shapes. Then, we find out that lips shapes (without buccal cavity, teeth, and a tongue and etc.) of opened mouth during speech have the same topological properties. In fact, they are homeomorphic. Thus, we use a new template shape illustrated in Figure 1 to cope with the ever-changing lips region. This shape is called "square annulus".

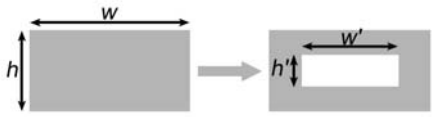

Figure 1. Square annulus

In Figure 1, w and h are the source square template's height and width. And w' and h' are the new square annulus template's width and height. In simulations (refer to Section 5), w' and h' are decided experientially. We performed preliminary examination to use this new template shape. In case of using the normal square template, the extraction accuracy rate is 10 [%], and in case of using the new square annulus template is 100 [%]. Thus, considering topology and using the square annulus is effective for lips extraction. And, the ignored $w' \times h'$ region reduces the accuracy is amount of calculated and makes the lips extraction high speed.

3.2. Control of Aspect Ratio

The aspect ratio of the template that can be transformed in matching processes such as scaling and rotation, is limited. The reason is, in some preliminary examinations, there are some failure that an extracted lips region is extracted using an impossible aspect ratio. This problem is caused by the redness of image data, the angle of light and the blushed part of face.

Therefore, we checked the aspect of varying lips at the moment of speech. The possible aspect ratio's range is shown in equation (1).

$$1.0 \leq \frac{w}{h} \leq 2.0 \tag{1}$$

where w and h are source square template's height and width, same values as in Figure 1.

The failure about the aspect ratio is eliminated by using this control.

4. Genetic Algorithm

4.1. Structure of Chromosome

A chromosome is an optimized solution. In other words, chromosomes are parameters which represent coordinates, scaling and rotation of an exploration object on the target image.

The solutions obtained by manual operations, is called true solution. Our method results are judged to be good or not good by comparison with true solutions. The comparison is performed by the following equations.

$$\begin{cases} C - 3 \leq c \leq C + 3 \\ M \leq m \leq 1.3 \times M \\ ANGLE - 5^\circ \leq angle \leq ANGLE + 5^\circ \end{cases} \tag{2}$$

where: capital letters are solutions obtained manually, and small letters are solutions by proposed method. c represents the x or y coordinate, m is scaling rate and $angle$ is rotation angle.

If a result satisfies these conditions, a good result for the speech recognition is obtained.

The Structure of Chromosome

Figure 2 shows the structure of chromosome before improvement. In Figure 2, c_x and c_y are coordinates after parallel translation, m_x and m_y are scaling rates, and $angle$ is rotation angle of lips shape. Each gene lengths are 8 bits and therefore, total chromosome length is 40 bits.

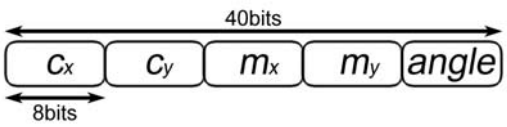

Figure 2. The new structure of chromosome

The lips region at the moment of speech is longer than the width or longer than the height. Thus, we use 2-dimensional scaling by m_x and m_y.

4.2. Fitness Function

In the SF-NPD method , we use only a fitness function which is normalised in range $[0, 1]$, because of reducing a dispersion of pixel differences whose pixel values are real values, x component of Yxy colour space. Moreover, this normalisation makes genetic algorithm process a simple maximisation problem. Against that, the DF-PD method does not normalise the pixel difference, and has an objective function and a dynamic fitness function. The objective function is a minimisation problem and the dynamic fitness function is a maximisation problem.

Pixel Difference

At first, parameters of geometric transformations are obtained from the chromosome. Then, the pixel difference is calculated by using equation (3)

$$D_{ij} = \begin{cases} \left|p^*_{ij} - p_{ij}\right| & \left(p^*_{ij} \in \text{target image}\right) \\ P_{\max} & \left(p^*_{ij} \notin \text{target image}\right) \end{cases} \quad (3)$$

where $P_{\max}$ is a maximum value of pixel, P is a point on the template image, and P^* is the point that corresponds to a transformed point P on the target image, p is a pixel value of a point P on coordinate (i, j) in the template image, p^* is a pixel value of a point P^* on coordinate (i, j) in the target image. D_{ij} is value of the pixel difference between p and p^*, however, in case of a point P^* is out of region template image, D_{ij} is worst $P_{\max}$.

Penalty

From preliminary examinations, we knew that the search efficiency is not very good and the result is unstable by using the fitness function mentioned below [1]. Therefore, the small pixel difference should not be ignored and the individual difference should be emphasised by introducing penalty for the pixel difference D_{ij}. Concretely speaking, if a pixel difference satisfies equation (4), D_{ij} is worst maximum value $P_{\max}$.

if

$$D_{ij} \geq TH \quad (4)$$

then

$$D_{ij} = P_{\max} \quad (5)$$

where TH is the threshold value to penalize.

This value is calculated by the following equation:

$$TH = \left(1 - \frac{S}{100}\right) \times P_{\max} \quad (6)$$

where S is the similarity per pixel, decided experientially and the unit is $[\%]$.

The more S increases, the stricter a condition becomes. In simulations, S is 90. Because, a probability of $100[\%]$ match is very low.

SF-NPD Method

A fitness value is calculated by using equation (7). The fitness is normalised and static.

$$fitness = 1.0 - \frac{\sum_{j=1}^{h}\sum_{i=1}^{w} D_{ij}}{(w \times h)\,(P_{\max})} \quad (7)$$

where the template size is w and h.

$fitness$ allows us to do a good exploration that the fitness approaches to 1.

DF-PD Method

An objective value and a fitness value is calculated by using equation (8) and (9). The fitness is not normalised and dynamic.

$$O = \sum_{j=1}^{h}\sum_{i=1}^{w} D_{ij} \quad (8)$$

where O is value of the objective function.

$$fitness = \max\{W_t, W_{t-1}, \ldots, W_{t-n}\} - O \quad (9)$$

where t is a current number of generation, W_t is a worst objective value of generation t, and n is a number of generation before.

This fitness function changes dynamically according to the worst objective value during n generations. Therefore, the selection pressure can be controlled automatically. O allows us to achieve good exploration that the value of the objective function approaches 0. In other words, $fitness$ allows us to do a good exploration that the fitness value becomes large. We use $n = 5$ in simulations.

5. Results and Discussions

5.1. Input Images

The template images are illustrated in Figure 3. Template image size of subject 1 and subject 2 is 18×8 [pixels], and subject 3 is 20×9 [pixels].

Figure 4 below shows examples (pronounce the vowel /e/) of target images. The images were captured using a

video camera include a face and background while each of three objects pronounce the vowels. Target images are then cut from the video streams. In consideration of use by mobile devices, target images have geometric changes based on the template image. The geometric changes are such as parallel translation, scaling, rotation. These value can be regarded as the true solution (refer to 4.1). All target images size are 240×180 [pixels].

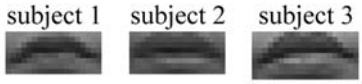

Figure 3. Template images

Figure 4. Target images

5.2. Configurations of System

A new structure chromosome is used in this demonstration (refer to 4.1). The parameters of genetic algorithms are: population size is 70, probability of crossover is 0.7, and probability of mutation is 0.05. The shape of template change parameters is $w'/w = 0.8$, $h'/h = 0.5$ (refer to Figure 1 in 3.1). Parameter S which decides a threshold of penalizing (refer to equations (4), (5) and (6) in 4.2) is set to 90. We use $n = 5$ in equation (9) of the DF-PD method. If the same fitness value continues until some generations, the solution is regarded as having converged and extraction is terminated. The more this value becomes large, the more the termination criterion becomes fair.

The machine speck which we use for simulation is Pentium4: 2[GHz].

5.3. Result of Simulation and Consideration

Figure 5 shows examples of results from the computer simulation. The rectangle region is the extracted lips region. The shape deformations of lips by speech are extracted exactly as shown in Figure 5.

Figure 5. Result images

Table 1 shows the true solution obtained manually for /e/ of subject 2 in Figure 4 and the solution obtained by the proposed method. It is found that these both solutions are similar.

Table 1. Example of solution of result (subject 2 /e/)

	coordinate		scaling [rate]		rotation
	x	y	x	y	[deg]
true solution	82	128	1.39	2.00	19.20
experimental result	81	131	1.68	2.51	17.33

The effectiveness of our method is demonstrated using 20 times simulations for each person (total is 60 times simulations per one vowel) being tested as shown in Tables 2, 3 and 4.

Table 2. Results of simulation (DF-PD method, tough criterion)

	/a/	/i/	/u/	/e/	/o/	total
acc [%]	95.00	93.33	95.00	91.67	98.33	94.67
time [sec]	0.198	0.203	0.199	0.197	0.199	0.199
gene	78.48	81.15	80.06	77.25	76.50	78.69

Table 3. Results of simulation (SF-NPD method, tough criterion)

	/a/	/i/	/u/	/e/	/o/	total
acc [%]	26.67	31.67	31.67	33.33	18.33	28.33
time [sec]	0.065	0.069	0.084	0.069	0.065	0.071
gene	55.52	63.24	75.41	60.84	57.42	62.49

Table 4. Results of simulation (SF-NPD method, fair criterion)

	/a/	/i/	/u/	/e/	/o/	total
acc [%]	98.33	96.67	91.67	93.33	91.67	94.33
time [sec]	0.323	0.318	0.332	0.299	0.329	0.320
gene	140.7	139.9	147.1	130.8	144.2	140.5

In Tables 2 and 3, the configuration of the simulations are described in 5.2. Table 4 is tested with another configuration and shows the conventional method result which is described in [2]. In case of Tables 2 and 3, their termination criterion value is 25 generations (refer to 5.2). Against that, in case of Table 4, it is 50 generations. In other words, Table 4 criterion is fairer than that in Tables 2 and 3.

In Tables 2, 3 and 4, "acc" means "extraction accuracy (only for correctly extracted lips region)", "time" means "processing time" and "gene" means "generation". Additionally, in Table 3, processing time and generation are not important, because their extraction accuracies are very low.

Comparing Table 3 with Table 4, they indicate that the SF-NPD method obtain high extraction accuracy on the fair

criterion, however on the tough criterion, it obtains low extraction accuracy. Against that, the DF-PD method works on the tough criterion in Table 2, and obtains better accuracy and speed.

Figures 6 and 7 illustrate a typical transition of the elite fitness and objective value with the SF-NPD method and the DF-PD method. Good results such as Figure 5 are obtained by these simulations. In Figure 6, the elite fitness changes with very small value and narrow range as opposed to Figure 7. This indicated that Figure 7 performed more efficient exploration. The reason is that the DF-PD method controls the selection pressure dynamically by equation (9).

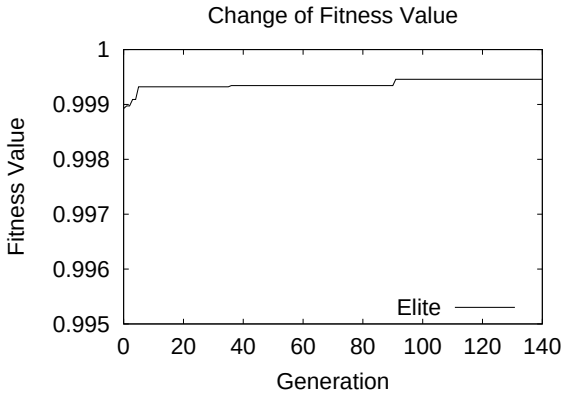

Figure 6. Transition of elite fitness (SF-NPD)

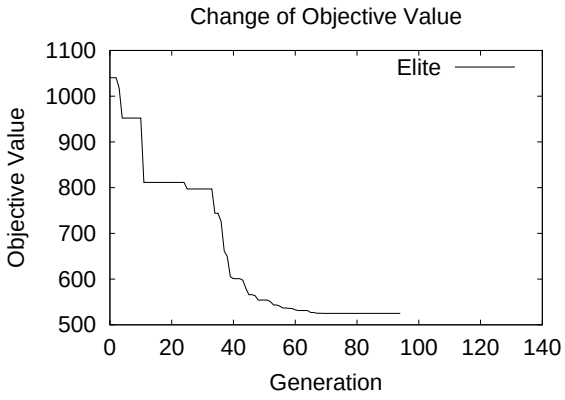

Figure 7. Transition of elite objective value (DF-PD)

6. Conclusion and Future Work

In this paper, a method for the improvement of lips extraction that has invariance for an opened and closed mouth showing or not showing any teeth, and has high speed and high recognition accuracy using only one template, is proposed. In Section 5, the results of these computer simulations indicate that the DF-PD method can perform more efficient exploration and has invariance for the varying shape, and higher extraction accuracy and a higher speed can be obtained in the extraction processing of all the vowels.

However, in the DF-PD method, some parameters have a lot to do with experiences. For example, S which decides the threshold of penalizing is set to 90 by trial and error (refer to 4.2), and the unconsidered region size w', h' of the square annulus is fixed (refer to 5.2). Moreover, parameter n of equation (9) in the DF-PD method is fixed to 5. Therefore, our future work needs improvement to decide all parameters automatically. Other future work is to apply this method to real-time processing. For this improvement, we hope to consider use of other feature of lips, such as another shape of the template except for the square annulus, will be effective.

References

[1] T. Akashi, M. Fukumi, and N. Akamatsu. Lips extraction with template matching by genetic algorithm. In *Knowledge-Based Intelligent Information Engineering systems and Allied Technologies*, pages 343–347, Crema, Italy, September 2002.

[2] T. Akashi, Y. Mitsukura, M. Fukumi, and N. Akamatsu. Genetic lips extraction method for varying shape. In *Computational Intelligence in Robotics and Automation*, Kobe, Japan, July 2003. (to appear).

[3] A. Hara and T. Nagao. Extraction of facial region of arbitrary directions from still images with a genetic algorithm. *Technical Report of IEICE*, HCS97-12, 97(262):37–44, December 1997.

[4] H. Kashiwazaki and S. Hirose. Extraction of issue for spoken language learning -using perceptual decision of spoken language and sound spectrograph-. In *The 17th Annual Meeting of the Japanese Cognitive Science Society*, pages 1–35, Japan, June 2000.

[5] M. Konda, M. Nishida, M. Ishii, and K. Sato. Application of the feature of lips motion in continuous image to personal identification. *T. IEE Japan*, 120-C(5):765–766, May 2000.

[6] S. Masunaga and T. Nagao. Extraction of human facial regions in still images using a genetic algorithm. *Technical Report of IEICE*, PRU95-160, 95(365):13–18, November 1995.

[7] L. Minolta Co. Story of color knowledge, 2002.

[8] Y. Nakata and M. Ando. Detection of mouth position using color extraction method and eigentemplate technique for lipreading. *Technical Report of IEICE*, PRMU2001-09, 101(303):7–12, September 2001.

[9] K. N. Plataniotis and A. N. Venetsanopoulos. *Color Image Processing and Applications*. Springer-Verlag, New York, USA, 2000.

[10] F. Saitoh. Pose recognition of gray-scaled template image using genetic algorithm. *T. IEE Japan*, 121-C(10):1500–1507, October 2001.

[11] K. Tanaka, M. Sano, S. Ohara, and M. Okudaira. Parametric template method and its application to high-accuracy robust matching. *T. IEICE Japan*, J83-D-II(4):1119–1130, April 2000.