High Speed Genetic Lips Detection by Dynamic Search Domain Control

Takuya Akashi Member (Yamaguchi University, akashi@eee.yamaguchi-u.ac.jp)
Yuji Wakasa Member (Yamaguchi University, wakasa@eee.yamaguchi-u.ac.jp)
Kanya Tanaka Senior Member (Yamaguchi University, ktanaka@eee.yamaguchi-u.ac.jp)
Stephen Karungaru Non-member (The University of Tokushima, fukumi@is.tokuchima-u.ac.jp)
Minoru Fukumi Member (The University of Tokushima, karunga@is.tokuchima-u.ac.jp)

Keywords: Genetic algorithm, Template matching, Lips image, Scaling window, Dynamic search domain control

In this paper, high-speed size and orientation invariant lips detection of a talking person in an active scene using template matching and genetic algorithms is proposed (refer to Fig. 1). As part of the objectives, we also try to acquire numerical parameters to represent the lips. The information is very important for many applications, where high performance is required, such as audio-visual speech recognition, and personal mobile devices interfaces. The difficulty in lips detection is mainly due to deformations and geometric changes of the lips during speech and the active scene by free camera motion. In order to enhance the performance in speed and accuracy, initially, the performance is improved on a single still image, that is, the base of video processing.

Our proposed system is based on template matching using genetic algorithms (GA). Only one template is prepared per experiment. The template is the closed mouth of a subject, because the application is for personal devices. In our previous study, the main problem was trade-off between search accuracy and search speed. To overcome this problem, we use two methods: scaling window and dynamic search domain control (SD-Control).

The transition of the objective value on the proposed system and the no SD-Control are plotted in Fig. 2. At points from (a) to (d) in Figs. 2, the transition of GA exploration is exactly the same in both cases. The reason is that the same random numbers are used in both cases. Between points (d) and (e) in Fig. 2, the differences in the methods appear. This is because, the search domain is controlled from point (d) in Fig. 2. The elite individual of the no SD-Control do not change from point (c) to (g), hence it is concluded that the GA exploration gets trapped into a local optimum. The reason why the exploration was trapped into a local optimum is that the small population loses the population diversity and causes premature convergence.

On the other hand, in case of the proposed system, the evolution make good progress from generation = 117 (point from (e) to (i)). This indicates that the GA of the proposed system can escape from the local maximum. The reason for this is that a search domain is reduced by the dynamic SD-Control and the GA can explore the reduced search domain with meticulous detail. This means that the GA in the proposed system performs not only global optimization but also local optimization. The result of our demonstration clearly shows that a trade-off between exploration accuracy and speed is overcome by the dynamic SD-Control.

We achieved a lips detection accuracy of 91.33% at an average processing time of 33.70 milliseconds per frame.

Fig. 1. Examples of image: a) target; b) template; c) detected lips as steady-state

Fig. 2. Transition of objective value for the points from (a) to (i), where the generation is 0, 10, 43, 90, 120, 130, 140, 155, and 199.
High Speed Genetic Lips Detection by Dynamic Search Domain Control

Takuya Akashi∗  Member
Yuji Wakasa∗  Member
Kanya Tanaka∗  Senior Member
Stephen Karungaru∗∗  Non-member
Minoru Fukumi∗∗  Member

In this paper, high-speed size and orientation invariant lips detection of a talking person in an active scene using template matching and genetic algorithms is proposed. As part of the objectives, we also try to acquire numerical parameters to represent the lips. The information is very important for many applications, where high performance is required, such as audio-visual speech recognition, speaker identification systems, robot perception and personal mobile devices interfaces. The difficulty in lips detection is mainly due to deformations and geometric changes of the lips during speech and the active scene by free camera motion. In order to enhance the performance in speed and accuracy, initially, the performance is improved on a single still image, that is, the base of video processing. Our proposed system is based on template matching using genetic algorithms (GA). Only one template is prepared per experiment. The template is the closed mouth of a subject, because the application is for personal devices. In our previous study, the main problem was trade-off between search accuracy and search speed. To overcome this problem, we use two methods: scaling window and dynamic search domain control (SD-Control). We therefore focus on the population size of the GA, because it has a direct effect on search accuracy and speed. The effectiveness of the proposed system is demonstrated by performing computer simulations. We achieved a lips detection accuracy of 91.33% at an average processing time of 33.70 milliseconds per frame.

Keywords: Genetic algorithm, Template matching, Lips image, Scaling window, Dynamic search domain control

1. Introduction

Speech recognition is one of the most useful interfaces that requires little space and has no physical contact between the device and the user. However, in the mobile devices, it has a performance limitation due to background noise in various locations, such as public areas, stores, offices and home. In order to overcome this problem, lips information should be effectively used. In human speech perception, audio-visual integration is very useful (1)–(3). Therefore, speech recognition using the lips image is very important as an interface in mobile devices. Many studies of audio-visual speech recognition using lips image have been reported (4)–(9).

An image-based surveillance system with mobile robots, such as wheeled robot (10) and micro aerial robot (11) (12) has been proposed. Moreover, speaker identification by using lips information is performed in (13)(14). If these two technologies are collaborated, a human surveillance system with a mobile robot can be achieved, as mentioned in (15). Therefore, the lips are very important for interfaces of the personal mobile devices.

Moreover, the lips redness is common in the entire human race. The reason is that the redness is composed of non-keratinized squamous epithelium that covers numerous capillaries, which give the lips its characteristic color (16)–(19). For this reasons, we focus on the lips redness as a main feature during detection.

The purpose of our study is the lips detection and lips information acquisition as the image-based front-end of audio-visual speech recognition and speaker identification in a personal mobile device. In order to make the speech recognition and speaker identification processes simple and easy, it is preferred that lips used in the process be of steady-state, as shown in Fig. 1(c).

During the usage of mobile devices with cameras installed, the camera and the user movements are usually independent of each other. Therefore, geometric changes
in the lips, such as parallel translation, scaling and rotation must be corrected. For a real-time application, the system should process both the orientation of the lips and acquisition of lips information simultaneously.

The lips have two types of features; shape and anatomical color as mentioned above. Both these features should be used during detection. In order to achieve our objective, we use single template matching using a genetic algorithm during the matching process. At first, input images, that is, a template image and target images are prepared. Normally, when a template is being created, the changes in subjects, scene, and lips must be considered. However, in this study, we focus only in the lips changes by speech and camera motion because this system will be applied to personal devices. It is difficult and time-consuming to create many templates of the various changes of the lips. However, in this work, only one template is created. An example of the template is illustrated in Fig. 2.

The template image is acquired from a face image captured just before video sequence capturing for target images. In this system, the basic shape of the template is a square meaning that both the skin color and the lips color are contained, as shown in Fig. 2(b). From anatomical viewpoint, color difference between skin and lips is a very important feature that makes lips shape calculation easy. The template image is a closed mouth of the same subject in the target images because the application of this system is for personal devices. Next, the source video sequence is captured while the subject is talking using free camera motion Fig. 3. The lips in the target image have various shapes as shown in Fig.3(b). The types of lips shapes are described in Section 3.

In our previous report, we used single template matching using GAs for one frame from a video sequence. The GA’s chromosome in the system specified geometric information of lips. Therefore, we could detect the lips region, even in cases where the lips region had some significant geometric changes by free camera motion. However, there is a trade-off between accuracy and a processing time in the system that makes it impossible to use the system in real-time processing. In the the previous study, although the accuracy was high 90.50%, the processing time was slow 273.90 milliseconds. Consequently, we propose a new method; dynamic search domain control (hereinafter referred to as SD-Control) that can make the system applicable in a real-time processing.

As a result, the use of the dynamic SD-Control was very effective and the performance of the proposed method was better than before.

From results of the demonstrations, we can conclude that the proposed method can overcome the trade-off problem between exploration accuracy and speed, an ascribable nature of GA.

The paper is organized as follows: the problem which is treated in this study, is described in the next Section 2. Section 3 presents an overview of the lips information which must be acquired, and the template shape. In Section 4, our techniques are explained. Section 5 shows an evaluation of the system by experiments and comparisons with the reference. Section 6 describes future works. Finally, Section 7 gives a conclusion.

2. Statement of the Problem

This section clarifies our objectives in this work and also describes some of the works related to it.

The purpose of our study is the detection of a lips region and the acquisition of the lips information as an interface and front-end of the audio-visual speech recognition on personal mobile devices. Hence the camera and the user move independently and the lips region can have some significant geometric changes, such as parallel translation, scaling and rotation. Moreover, the lips shape changes during speech.

In this paper, we address three issues for lips detection:

1. Active scene by free camera motion.
2. High accuracy in detection of lips region and acquisition of lips geometric information.
3. High processing speed.

2.1 Related Work Many studies of audio-visual speech recognition by lips image have been reported. As far as we know, most of these studies set a precondition for the lips image. For example, in the subject wears a helmet with a camera and in lips images extracted beforehand from a database are used. Moreover, in the lips information is acquired through the lips detection after detection of a static face. In real-time processing, lips detection and lips information acquisition must be performed in only one phase. Image processing methods are divided into three main approaches; an image-based approach, a model-based...
approach and a combination of the two, an image- and model-based approach. In the image-based approach, eigenlip methods (7) have been proposed. In these methods, a set of training lips images is generated by the principal component analysis. The training data must be chosen carefully to include all possible lips configurations. Other image-based approaches proposed are rule-based approaches by features of a face (23), pixel-based approaches using red exclusion (25), optical-flow approaches to measure lips movement (26), etc. However, these approaches cannot adapt to considerable geometric changes in lips in every frame. In the model-based approach, active shape model (26) and genetic snakes (8), that is, an improved version of Snakes (27), have been proposed. These approaches have some constraints, such as, the target image must be a face region of fixed pose and the number of nodes and parameters should be skillfully determined. Therefore, it is difficult to apply these approaches to meet our objective. As for image- and model-based approaches, high-speed face tracking method (27) is proposed. In this method, many facial feature patch templates must be prepared as a training set. These templates are regions surrounding the feature, such as eye and mouth. Therefore, the geometric information of lips region cannot be acquired during the detection. A method using the whole face would also be difficult to use in our work because simultaneous lips detection and information acquisition should be accomplished in real time.

According to the explanations in reference (20), we have proposed a genetic lips extraction method. This method uses one closed mouth template and genetic algorithms. In this reference, the results indicated that it took 273.90 milliseconds to obtain 90.5% accuracy. This previous method was not suitable for the real-time processing. Therefore, in this paper, we implement other fitness function and propose the dynamic SD-Control method to achieve high accuracy exploration and fast search.

2.2 Feature of Lips Color Since the lips redness is a common feature in the entire human race (Section 1), its data is useful for the matching process. Therefore, the source color data (non-linear RGB data) is converted into the CIE-Yxy (28) color space’s x component. The CIE-Yxy color space is illustrated in the chromaticity diagram, Fig. 4. The x component represents redness as shown in Fig. 4.

Moreover, lightness does not greatly influence the x component, because in the CIE-Yxy color space, the components are presented independent of lightness (28). Therefore, because of these merits, the x component is used in this study. Comparing the modified x component with other color space values in our preliminary experiment, it achieved the best result. Fig. 5 shows the conversion from the source data to the x component data.

3. Lips Information and Template Shape

This section describes the lips information that must be acquired and the shape of the template. At first, the changes of lips are classified according to their causes. Next, the relation between the classification and geometry are described. Taking into account this relationship, a template shape is explained that is appropriate for the lips detection.

3.1 Change of Lips and Transformation Group

Depending on the cause, the lips shape changes can be divided into two types. One, changes of the lips region by the independent motion of a camera and user. In this case, the lips region has changes, such as parallel translation, scaling, and rotation. Two, lips shape deformation by speech.

Before description of the relation between the lips changes and geometry, an overview of the types of geometric transformation (29) possible is discussed. The hierarchy structure of transformation is illustrated in Fig. 6. In the figure, “G” and “L” are considered as graphic symbols. The base shape is Fig. 6(a). The innermost group is the Euclidean transformation; Fig. 6(b) is a result of rotation. Following the Euclidean transformation group is the similarity transformation group; Fig. 6(c) is a result of scaling. After that is the affine transformation group, Fig. 6(d) is a result of shear and reflection. The outermost group is the topological transformation. The connectivity of all “G” and “L” are the same. Therefore these relations are homeomorphic.

The affine transformation group can represent the lips region changes caused by motion of the camera and the
user. This is because this lips region change has geometric changes, such as parallel translation, scaling, and rotation.

When considering the change of lips shape by speech, in most cases, the mouth is open during speech. Assuming that the lips deform like an elastic band, during speech the lips can be considered to expand and contract like the band. In other word, its connectivity is constantly same. Therefore, the topological transformation group can represent this change. It then follows that the various transformation groups discussed can represent all the possible lips region changes.

3.2 Shape of Template In general, the typical template shape is a complete square. This shape is suitable for geometric changes, such as parallel translation, scaling, and rotation. However, considering an application of template matching to the change of lips shapes during speech, the complete square shape of template is unsuitable. This is because, at the moment of speech, the lips region has intense variations such as an open or closed mouth and showing or not showing any teeth. For this reason, we have used a template shape “square annulus” (20) that considers the lips changes and their relation to the transformation group as mentioned above (Section 3.1). This template shape is illustrated in Fig. 7.

In Fig. 7, \( w \) and \( h \) are the source square template’s width and height respectively. \( w' \) and \( h' \) are width and height of an ignored region in the “square annulus” respectively. In our preliminary examination using “square annulus”, the detection accuracy went up considerably, compared with the normal square template. Furthermore, another advantage of the “square annulus” is that the \( w' \times h' \) region ignored during matching reduces the amount of calculation, further improving the speed of lips region detection. In this work, in order to simplify the system, we fix the interior area dimensions; the interior height \( h' \) and width \( w' \) are set to 80% and 50% of the exterior height \( h \) and width \( w \) respectively. These values are decided by simulation results in the reference (20). The reason why the same value as in the reference (20) is used is that they are compared under the same condition to evaluate the effectiveness of the proposed method.

4. Genetic Lips Detection

In this section, the details of our proposed lips detection and lips information acquisition system are described as follows: structure of chromosome in GA, fitness function, dynamic SD-Control and a flow chart of the proposed system is also provided.

4.1 Structure of Chromosome In GAs, an individual is a solution candidate to be optimized. The individual consists of chromosomes that are the source of the solution. In our optimization problem, this solution represents a transformation matrix that transforms the template on the target image. In fact, the template is transformed by the matrix in homogeneous coordinates before template matching is performed.

The structure of the chromosome is shown in Fig. 8, where \( t_x \) and \( t_y \) are coordinates after parallel translation, \( m_x \) and \( m_y \) are scaling rates, and \( \text{angle} \) is rotation angle of lips region. These parameters are called phenotype and are encoded in some bit-string, which is called a genotype. In our GA, the coordinates are coded in 8-bit for the range zero to the size of target image, the scaling rates for the range 0.8 to 3.0 are also coded using 8-bits. The rotation angle is likewise coded in 8-bit for the range -35 to 35 degrees. The total chromosome length is then 40 bits. The template’s width and height should be changed separately because the shape deformation of lips by speech is not confined to similarity changes only. Consequently, we use 2-dimensional scaling by \( m_x \) and \( m_y \).

4.2 Fitness Function In this part, a fitness function that evaluates the individuals is explained. Although a static fitness function is used in the reference (20), a dynamic fitness function is used in this paper. This reason is that the dynamic fitness function can control the selection pressure dynamically, and the performance is better than the static fitness function, as mentioned in a reference (21).

At first, the template is transformed on the target image by the chromosome as mentioned above. After the transformation, a pixel difference is calculated between the template and a target image. Then, an objective
function and a fitness function are calculated. The objective function is regarded as a minimization problem while the fitness function is regarded as a maximization problem.

4.2.1 Pixel Difference After the transformation of the template on the target image by the chromosome, the pixel difference between the template and target image is calculated as follows:

\[
D_{ij} = \begin{cases} 
  a_{ij}^* - a_{ij} & \text{if } (a_{ij}^* \in \text{target image}) \\
  A_{\text{max}} & \text{if } (a_{ij} \notin \text{target image}) 
\end{cases} \quad (1)
\]

where \( A_{\text{max}} \) is the maximum value of pixel, \( a \) is a pixel value of a point \( A \) on coordinate \((i, j)\) in the template image, \( a^* \) is a pixel value of a point \( A^* \) in the target image. This \( A^* \) is the point on the target image that corresponds to a point transformed from the point \( A \). \( D_{ij} \) is a value of the pixel difference between \( a \) and \( a^* \), however, in case that a point \( A^* \) is out of region in the template image, \( D_{ij} \) is the worst \( A_{\text{max}} \). In this system, the pixel value is the x component value as described in Section 2.2.

4.2.2 Objective Function and Fitness Function

An objective function and a fitness function are defined by Eq. (2) and (3) respectively. The fitness function is dynamic:

\[
O = \sum_{i=1}^{w} \sum_{j=1}^{h} D_{ij}, \ldots, \ldots, \ldots, \ldots \quad (2)
\]

\[
\text{fitness} = \max \{W_t, W_{t-1}, \ldots, W_{t-n}\} - O, \ldots, \ldots, (3)
\]

where \( O \) is the objective value, which is a summation of \( D_{ij} \). fitness is the fitness value, which is a difference between \( O \) and a worst objective value for the last \( n + 1 \) generations. \( t \) is a current index of generation. The fitness is the difference between the worst objective value encountered earlier and the current \( n \) generations. This technique is called “scaling window” \((20)\) \((21)\) is used for controlling selection pressure of GA. The control is very important because the selection pressure and population diversity are inversely related \((21)\). In other words, as the selection pressure is increased, the population diversity decreases usually causing a premature convergence. Conversely, the lack of selection pressure can cause evolutionary retardation. Another approach for controlling selection pressure is “ranking” \((22)\). However, the search speed is slow except in some special cases \((21)\). Therefore, this work uses the scaling window method.

In reference \((20)\), we used the static fitness function. We demonstrated the effectiveness of the dynamic fitness function with the scaling window by comparative experiments in reference \((21)\). The results indicated that the dynamic fitness function is effective for our optimization problem, however the decision of an important parameter, window size \( n \) is a problem. In general, if the size is too large, GA exploration depends on a longstanding worst individual slowing the GA search. Against that, if the size is too small, GA exploration is sensitive to noise and to an incidental good individual hence GA exploration can easily be trapped in a local optimum. To the best of the authors’ knowledge, no investigation of this parameter has yet been carried out. In this paper, we investigate the influence of \( n \) on results in Section 5.3.

4.3 Dynamic Search Domain Control

The method proposed may be unsuitable for the real-time video sequence, because of slow performance. The slow performance is due to the trade-off between exploration accuracy and speed.

The computational cost of one individual in GA is calculated in two steps as follows.

1) Generate the next generation: selection, crossover, and mutation.

2) Calculation of a fitness value: de-coding of the chromosome to the phenotype, transformation of the template, and calculation of the pixel difference one by one pixel.

Moreover, the computational cost of one generation is directly proportional to the size of population (the number of individual in a generation). The size of the population and the number of generations must be increased to obtain high accuracy. This in turn reduces the search speed of the GA. From earlier experiences, when the size of the population and the number of generations are decreased, GA individuals can be stuck in local optima, Fig. 9. This is because the GA is a global optimization algorithm and is not good for local optimization \((22)\).

In GAs, the search starts from a population of many points, rather than starting from just a single point. This parallelism means that the search will not become trapped in local optima. GA tries to escape from the local optimum and to find the global optimum by crossover and mutation operators. If the population is too small, GA converges prematurely and is trapped in a local optimum. As a search efficiency improvement, we can use a technique, in which after the domain including the optimal solution is specified, its neighborhood is searched in detail.

By controlling the search domain of GA, we expect that it will become easier for the GA to escape from the local maximum and that the GA can be used not only for global optimization but also for local optimization. We thus hope that by decreasing the population and controlling the search domain, high accuracy exploration and fast search can be achieved.

The search domain is controlled depending on both an elite individual and the number of generations. The elite individual can be found out by comparison of the objective value of all individuals. The location of the search domain is decided by a coordinate \((t_x \text{ and } t_y, \text{see Fig. 8})\) of the elite individual. The search domain center is set to this coordinate.

Next, the size of the search domain is decided by the

Fig. 9. Examples of the local optimum
number of generations. The search domain is renewed as follows:

\[
\begin{align*}
\text{width}^* & = \alpha \text{width} \\
\text{height}^* & = \alpha \text{height}
\end{align*}
\] (4)

In Eq. (4), width and height are the target image's width and height, transformed to width* and height* respectively by \(\alpha\) which is a scale factor. This \(\alpha\) is controlled by the number of generations. Initially, the detection in the early stage of the GA evolution is very important. Therefore, the search domain is full range for generation < 90. The value of \(\alpha\) can be defined as follow:

\[
\alpha = \begin{cases} 
1 & (\text{generation} < 90) \\
0.9 & (90 \leq \text{generation} < 100) \\
0.8 & (100 \leq \text{generation} < 120) \\
0.7 & (120 \leq \text{generation} < 130) \\
0.6 & (130 \leq \text{generation} < 140) \\
0.5 & (140 \leq \text{generation})
\end{cases}
\] (5)

where \(\text{generation}\) is the number of generations. In Eq. (5), the search domain is reduced in multi-step as evolution progresses. This means that the search domain is controlled dynamically (dynamic SD-Control). Although there are many combinations of the value of \(\alpha\) and the range of \(\text{generation}\), we use Eq. (5) in this paper. This reason is that the lips detection accuracy with Eq. (5) is best in some exploratory experiments.

Some individuals can be located outside of the search domain during the search region change using the SD-Control. However, individuals cannot be eliminated because all genes may evolve into good direction except genes that represent the position. Therefore, whenever the search domain is changed, only \(t_s\) and \(t_p\) in the chromosome is re-coded for all individuals by a new search domain. Using this process, other genes are inherited to the next generation. This means that the GA optimization is controlled.

Generally speaking, this is a risky method, because it is not guaranteed that the domain where the optimal solution is included clearly, can be specified. In other words, this method can not find the optimal solution because of premature convergence to local optima. However, this is not a critical problem for our system. Typically, local optima in our simulations are a part of the face (see Fig. 9) because, skin area is included in the template image and redness data (refer to Section 2.2) is used. This means that it is highly possible that the optimal solution is in the neighborhood of local optima.

There is a generic computer vision technique in which the search domain is restricted by comparing local color histograms between the template and target image. This method can be used before GA processing. However, it is difficult to decision of a threshold, which restricts the search domain. Moreover, if the similar color with the lips includes the background, the search domain will expand or separate into some places. Therefore, we use the dynamic SD-Control in the proposed system.

4.4 Flow Chart

Flow charts of our system are illustrated in Fig. 10.

At first, an initial population is generated and after that the GA process is started. In GA processing, the template shape is deformed to an unique “square annulus” from a normal square, as explained in Section 3.2. Then, the matching process is executed between the template and a target image using the fitness function. The generation is increased, until a termination condition of GA is satisfied. In this paper, the GA is terminated by the number of generations. If the termination criterion is not satisfied, a new population of the next generation is generated according to the fitness of each individual (Fig. 10(b)). In this process, the search domain is controlled dynamically, as described in Section 4.3. This technique is the main part of this proposed system, and achieves high speed and the high accuracy lips detection. After GA process is completed, the result is obtained as numerical data. This numerical data represents the lips information and can be used in many applications as described in Section 1.

5. Computer Simulation Results and Discussion

In this section, computer simulations to evaluate the effectiveness of the proposed system and discussions are described. First, we survey a relation between the scaling window size (refer to Section 4.2) and accuracy of the lips detection by various sizes, so that the scaling window size can be decided. Furthermore, based on the result, we estimate the effectiveness of this system by comparing it to the system without the dynamic SD-Control (hereafter referred to as “the no SD-Control”). All simulations are performed on the same computer: CPU is Pentium4 2.0GHz.

5.1 Input Images for the Simulations

The input images in this paper are same images with the reference\(^{20}\). This reason is that there is not an image database which is suitable for our purpose as far as we
know. Moreover, using same images, the reference can be compared with the proposed method.

The template images are illustrated in Fig. 11, two images are prepared for each subject, one indoor and the other outdoors. The template image size of subject 1, indoor and outdoor is 20 × 11 and 21 × 13 pixels respectively. For subject 2, it is 23 × 11 and 24 × 10 pixels and for subject 3, the sizes are 22 × 11 and 22 × 12 pixels respectively.

The target image examples from video sequence taken by a digital video camera are presented in Fig. 12. Some red color objects are included in the background, for example, red objects on a poster and some flowers. Also note that the background is not controlled (it includes bicycles, books, etc.). During video capture, the subjects are asked to naturally pronounce the Japanese vowels, /a/, /i/, /u/, /e/ and /o/. Moreover, on the assumption that the camera moves and joggles by free hand, the shaking of scene is artificially done by hand. Therefore, the lips region has some geometric changes. All target images are 240 × 180 pixels in size. The alphabets in Fig. 11 and Fig. 12 are one-to-one correspondent, for example, the template Fig. 11(a) is used for the target images of Fig. 12(a).

5.2 GA Settings In this part, other GA settings are explained. As mentioned in Section 4.1, we use the binary genotype. For this encode, from the phenotype to genotype, we do not use binary-coding but Gray-coding, since Gray-coding is generally more superior. We choose uniform crossover, because of its many advantages. In this system, five parameters must be prepared as follows.

- Population size
- Crossover probability
- Mutation probability
- Scaling window size
- Termination criterion

These parameters affect the efficiency of GA exploration. The relation between the population size and search speed is described in Section 4.3. The crossover and mutation probabilities are decided after many trials. The scaling window is explained in Section 4.2.2. In Section 5.3, we survey a relation between the scaling window size and accuracy of GA exploration. This result is a basis of making decision on the size. For the GA termination, a variety of termination criterion has been used, such as the change of fitness value and the number of generation. In this system, we use the number of generations as described in Section 4.4, because of its simplicity and low computational cost. In Section 5.3, final values of these parameters used in simulations are shown for each simulation.

5.3 Experimental Result

5.3.1 Decision of Scaling Window Size As described in Section 4.2, the fitness function is changed dynamically using the scaling window. The scaling window size is one of the parameters listed in Section 5.2. Its size is important for the efficiency of GA exploration. However, as far as we know, there has been no report about the relation between scaling window size and accuracy of GA exploration. Therefore, in order to investigate the relation, we experiment using 32 scaling window sizes ranging from 0 to 30 and infinity, Table 1. The trial is conducted 20 times for each of 3 targets shown in Fig. 13. These target images are chosen from Fig. 12. Therefore, the total simulations done are 1,920 (32 × 3 × 20). Moreover, the dynamic SD-Control (refer to Section 4.3) is used.

Fig. 14 shows the relation between the window size and GA accuracy. Scaling window size 0 (n = 0 in Eq. (3)) means that the worst individual of current generation is used in the fitness function. Scaling window size ∞ means that the worst individual until current generation is used. In Fig. 14, the accuracy is very low at...
Table 1. Parameters in the simulation of scaling window size variety

<table>
<thead>
<tr>
<th>Population Size</th>
<th>Crossover probability</th>
<th>Mutation probability</th>
<th>Scaling window size</th>
<th>Termination criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>0.7</td>
<td>0.05</td>
<td>[0;30], ∞</td>
<td>100</td>
</tr>
</tbody>
</table>

These parameters are explained in Section 5.2.

Table 2. Parameters in the simulation of estimate the proposed system

<table>
<thead>
<tr>
<th>Population Size</th>
<th>Crossover probability</th>
<th>Mutation probability</th>
<th>Scaling window size</th>
<th>Termination criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.7</td>
<td>0.2</td>
<td>5</td>
<td>200</td>
</tr>
</tbody>
</table>

These parameters are explained in Section 5.2.

Fig. 13. Target images in simulations for relation between the scaling window size and accuracy. These are chosen from Fig. 12.

Fig. 14. The relation between scaling window size and GA accuracy.

scaling window size zero. The reason is that GA exploration is sensitive to noise and incidental good individual in a small scaling window size, dragging the exploration to a local optimum. At the other size infinity, the best size cannot be chosen from the result in Fig. 14. It is likely that there are two reasons. One is to trap into a local optimum due to the small scaling window size, and the other is slow convergence which depend on the past worst individual due to the big size window. As a result, we chose a value that is not too small, a scaling window size of 5 in the following demonstrations.

5.3.2 Demonstration of Search Domain Control

This systems main approach is the dynamic SD-Control method (refer to Section 4.3). Thus, the system effectiveness is estimated by comparison with the no SD-Control system.

Table 2 shows parameters of simulations. As argued in Section 4.3, in order to speed-up the GA exploration, the population size is very small, 10. The probability of crossover is standard, however the mutation probability is little high, because of the few individuals used. The size of scaling window is decided by the result of Section 5.3.1. The termination criterion is 200 generations. The test images used in the simulations are shown

Fig. 15. Examples of successful result image for Fig. 12 images

Fig. 11 and 12.

Fig. 15 shows successful result images for images in Fig. 12. In these results, the filled rectangle region is the detected lips region. The larger un-shaded rectangle represents the final search domain by the dynamic SD-Control. These results should however be verified manually.

As described in Section 4.1, this system can acquire the lips information directly. The system’s solutions are compared with manual solutions to verify the accuracy. Therefore, the difference between the manual and the system’s results gives the precision value.

The results of lips region detection are judged to be good or not good by comparison with the true solution. The comparison is performed using the following equations.

\[
\begin{align*}
T - 3 & \leq t \leq T + 3 \\
M & \leq m \leq 1.3 \times M \\
ANGLE - 5^\circ & \leq angle \leq ANGLE + 5^\circ
\end{align*}
\]

\[6\]
where these capital letters are the solution obtained manually, and small letters are solutions obtained by the proposed system. \( t \) represents the \( x \) or \( y \)-coordinate, \( m \) is a scaling rate and \( \text{angle} \) is a rotation angle.

If a result satisfies these conditions, the result is acceptable for the applications described in Section 1. In Eq. (6), if the system solution is smaller than actual lips size, the solution is rejected, because, a partial lips region cannot be used directly in the applications.

Table 3 shows lips data for Japanese vowel /a/ in Fig. 15. All system solutions of Table 3 satisfy the conditional equation Eq. (6). These results indicate that the proposed system can robustly and accurately detect lips region in images taken by free camera motion and deformed by speech.

For some applications described in Section 1, the lips can be corrected using the acquired lips data in Table 3. Examples of the corrected lips images are shown in Fig. 16. These images are generated by lips data from Table 3 for target images Fig. 12-/a/. Moreover, the heights of these corrected lips images are normalized. These corrected lips image can be used for many applications, such as audio-visual speech recognition, speaker identification system, robot perception and interface of personal mobile devices.

To compare this system with the no SD-Control system, simulations were carried out 20 times for each system using the 30 targets images in Fig. 12 (a total of 1200 (20 \( \times \) 30 \( \times \) 2) simulations). In the simulations, initial population and position of the crossover and mutation in the chromosome are given randomly. The same random value must be used to compare under the same conditions. Therefore, 600 different random seeds are prepared.

Tables 4 and 5 shows the comparison results. Additionally, in order to find the difference in Japanese vowel pronunciations the results of each vowel is described. As can be seen in Table 4, in both cases, the average processing time is almost the same. The reason is that, GA termination criterion is set to 200 generations (see Table 2). The processing time is close to the real-time processing because of the very small population size 10 and the termination criterion.

A comparison of the accuracies is shown in Table 5. If the system solution satisfies the conditional equation Eq. (6), the solution is correct. In Table 5, in the case of no SD-Control, the average of accuracy is 70.50%, while the proposed method accuracy is 91.33%. From these results, it can be said that the proposed method resolved the trade-off problem better than no SD-Control. In order to verify this, we compare a difference in the transition of the GA evolution between the two approaches.

The transition of GA evolution is shown in Figs. 17-19. The transition of the objective value (Section 4.2.2) on the proposed system and the no SD-Control are plotted in Fig. 19. The visual transition of GA evolution on the no SD-Control and the proposed system appear in Figs. 17 and 18 respectively (These cases are the most typical examples). At points from (a) to (d) in Figs. 19, 17, and 18, the transition of GA exploration is exactly the same in both cases. The reason is that the same random numbers are used in both cases. Between points (d) and (e) in Fig. 19, the differences in the methods ap-
Genetic Lips Detection by Dynamic Search Domain Control

This is because, the search domain is controlled from \( \text{generation} = 90 \) (point (d) in Fig. 19) as described in Eq. (5). The objective value of elite individual and visual transition of the no SD-Control do not change from point (c) to (g), hence it is concluded that the GA exploration gets trapped into a local optimum. Moreover, at point (h) in Fig. 19, the GA on the no SD-Control has a little evolution but the exploration was trapped into a local optimum again. Finally, the GA on the no SD-Control cannot reach a correct solution at point (i). The reason why the exploration was trapped into a local optimum is that the small population loses the population diversity and causes premature convergence.

On the other hand, in case of the proposed system, the evolution make good progress from \( \text{generation} = 117 \) (point from (e) to (i)). This indicates that the GA of the proposed system can escape from the local maximum. The reason for this is that a search domain is reduced by the dynamic SD-Control and the GA can explore the reduced search domain with meticulous detail. Fig. 18(e) shows that lips are detected, after that the GA fine-tunes the parameters, such as scaling and rotation angle. In other words, the GA in the proposed system performs not only global optimization but also local optimization. The result of our demonstration clearly shows that a trade-off between exploration accuracy and speed is overcome by the dynamic SD-Control.

6. Future Work

In this section, some problems that should be solved are described, and solutions for these problems are considered.

6.1 Future of Search Domain Control

In this part, some limitations and solutions of the proposed method are discussed.

6.1.1 Decision of Scale Factor and Timing of Changing

In this paper, we use Eq. (5) to change the search domain, because the result is best in some ex-

![Fig. 17. Visual transition of evolution by the no SD-Control system: these alphabets are correspondent with alphabets in Fig. 19.](image)

![Fig. 18. Visual transition of evolution by the proposed system: these alphabets are correspondent with alphabets in Fig. 19.](image)

![Fig. 19. Transition of objective value for the points from (a) to (i), where the generation is 0, 10, 43, 90, 120, 130, 140, 155, and 199.](image)
proposed method with Eq. 5 is better than reference (20). However, this is a problem for the complete automation of the proposed method. To solve this problem, the mechanism, which changes the search domain automatically, is necessary. For example, when evolutionary stasis of GA occurs, the method changes the search domain. We can know the condition of the evolution by observation of the fitness value. The evolutionary stasis means that GA is trapped in a local optimum. It is possible for the local optimum not to be the optimal solution. Therefore, it is useful to use the evolutionary stasis as a trigger to change the search domain.

6.1.2 Guarantee of Including the Optimal Solution Another problem is there is no guarantee of including the optimal solution in a search domain. To discuss this problem, we investigate the failure results of the proposed method. The accuracy is 91.3%, that is, the number of failures is 52 out of 600 trials, as shown in Table 5.

Features of the failed solution, which is found by the GA, as explained below. All failed solutions converge to a part of face (including a part of the lips: 10, other part such as nose or chin: 42), as shown in Fig. 20. This indicates that there are local optima on the face region. There are two reasons for this. First, in both, color information is similar. The reason of this similarity is that the template is acquired from a face image captured just before video sequence capturing for target images, as mentioned in Section 1 and 4.3. Therefore, the template contains skin color area. Another reason is the lack of the individual for the first processing speed.

The positional relation between the lips and last search domain, which is changed by the proposed method is explained below. The lips are included in the last search domain in 43 (7.2% of the total trials), and the lips are not included 9 (1.5% of the total trials). This means that the search domain does not include the optimal solution in certain random number. The reason for this are the rapid changes of the search domain size. To solve this problem, an efficient change of the search domain that does not miss the optimal solution, is necessary.

It can decrease the possibility of missing the optimal solution when the search domain is not only shrank but also enlarged.

Moreover, to change the shape of the search domain may be useful. In the proposed method, the search domain is always homothetic with the target image (horizontally long), however, the shape of the face is vertically long. Therefore, the possibility, that the search domain includes the lips, may be increased by changing the search domain to the vertically long shape. In the future, we will try these improvement methods on the basis of the proposed method.

6.2 Application for Audio-visual Speech Recognition The \( h' \) and \( w' \) represent the size of the ignored region in the “square annulus” in Fig. 7. These are important for the application to the audio-visual speech recognition, because these parameters represent how a mouse is opened. In this paper, these parameter are decided experimentally to compare with the reference (20), as mentioned in Section 3.2.

In our future plan, we will address how to estimate these parameters automatically for the speech recognition. For example, these parameter can be explored by GA, in other words, the \( h' \) and \( w' \) are added to the chromosome. This addition will make the exploration space expand, therefore, we will investigate the size of the population and the termination criterion, etc.

6.3 Other Improvements The color of lips is changed by the various illumination and the individual difference. The proposed system will be applied to the personal mobile device, thus we focus only in the lips changes by speech and camera motion. Therefore, the proposed system does not consider the changes of lips color. Taking into account the practical use and convenience of the proposed system in the future, the lips color changes should be supported. We will investigate the robustness of the proposed system to the lips color changes, and we will make the necessary improvements.

7. Conclusion

In this paper, we have shown the importance of real-time detection and information acquisition of lips of a talking person in active scenes with a personal mobile device. A high-speed lips region detection and data acquisition method using genetic algorithms with dynamic SD-Control was proposed.

Our previous system (20) has a trade-off problem between exploration accuracy and speed, an ascribable nature of GA. Although the accuracy was high 90.50%, the processing time was slow 273.90 milliseconds in the the previous study (20). In order to solve this problem, we proposed the SD-Control method and decided the scaling window size.

A size and orientation invariant lips detection accuracy of 91.33% in almost real time was achieved by the proposed system (the same compute was used in the reference (20) and this paper). As a result of simulations, it is concluded that this approach achieves high accuracy and high-speed detection using a very small population. Moreover, the proposed system can acquire the lips information at the detection of lips that as described, can be useful in many applications.

(Manuscript received Sep. 26, 2006, revised Jan. 11, 2007)
References


Takuya Akashi (Member) was born in Tokyo, Japan, on September 19, 1978. He received the B.E. in Information and Communication Science from Kyoto Sangyo University in 2001. He then re- ceived the M.E. degree and a PhD in the Sys- tem Design Engineering from the University of Tokushima in 2003 and 2006 respectively. Since April 2005, he has been with Depart- ment of Electrical and Electronic Engineering, Yamaguchi University. He is currently an as- sistant professor. His research interests include evolutionary algo- rithms, image processing and human sensing. He is a member of the IEEE, RISP and IEICE.

Yuji Wakasa (Member) was born in Japan in 1968. He re- ceived the B. S. and M. S. degrees in engineer- ing and the degree of doctor of informatics all from Kyoto University, Japan, in 1992, 1994 and 2000, respectively. He was a research asso- ciate at Okayama University and Kyoto Uni- versity, from 1994 to 1998 and from 1998 to 2002, respectively. Since October 2002, he has been with Department of Electrical and Elec- tronic Engineering, Yamaguchi University. He is currently an as- sistant professor. His current research interests include control system design via mathematical programming and its application. He is a member of SICE, ISICIE, ORSJ and IEEE.
Kanya Tanaka (Senior Member) was born in Japan, on 1952. He graduated from the Postgraduate Course in Science and Engineering of the National Defense Academy in 1982. From 1990 through 1992, he was an assistant professor of the Electronics Engineering at Hiroshima Denki Institute of Technology. From 1992 through 1998, he was an assistant professor of the Mechanical Engineering at Ehime University. He is currently a professor of the Information and Design Engineering at Yamaguchi University. His research interests include adaptive control and intelligent control. He is a member of the Society of Instrument Control Engineers, the Japan Society of Mechanical Engineers, and IEEE.

Stephen Karungaru (Non-member) was born in Kenya, on 1969. He received the B.Sc. degree in electronics and communication from Moi University in 1992. He then joined the Department of Electrical and Electronics Engineering, Jomo Kenyatta University of Agriculture and Technology as a teaching assistant. He received the M.E. degree and a PhD in information system design from the Department of information science and Intelligent Systems, University of Tokushima in 2001 and 2004 respectively. He is currently an assistant professor at the Institute of Technology and Science, The University of Tokushima. His research interests include pattern recognition, neural networks, evolutionary computation, image processing and robotics. He is a member of IEEE and ASM.

Minoru Fukumi (Member) was born in Tokushima, Japan, September 21, 1961. He received the B.E. and M.E. degrees from the University of Tokushima, in 1984 and 1987, respectively, and the doctor degree from Kyoto University in 1996. Since 1987, he has been with the Institute of Technology and Science, The University of Tokushima. In 2005, he became a Professor in the same department. He received the best paper award from the SICE in 1995 and best paper awards from some international conferences. His research interests include neural networks, evolutionary algorithms, image processing and human sensing. He is a member of the IEEE, SICE, ISCIE, IPSJ and IEICE.