Improving the Robustness of Lips Sensing with Evolutionary Video Processing

Takuya Akashi[1], Yuji Wakasa[1], Kanya Tanaka[1], and Minoru Fukumi[2]
[1]Graduate School of Science and Engineering, Yamaguchi University
2-16-1 Tokiwada, Ube, Yamaguchi, Japan
[2]Institute of Technology and Science, The University of Tokushima,
2-1 Minamijosanjima, Tokushima, Japan
E-mail : [1]{akashi, wakasa, ktanaka}@eee.yamaguchi-u.ac.jp, [2]fukumi@is.tokushima-u.ac.jp

Abstract: In this paper, an effective method is proposed for robust lips sensing. Our objectives are high-speed lips tracking and data acquisition of a talking person in natural scenes. Our approach is based on the Evolutionary Video Processing. This method has a trade-off between accuracy and a processing time. To solve this problem, we proposed automatic Search Domain Control method and implement this method in the Evolutionary Video Processing. The tracking accuracy is improved from 66.3% to 84.9%. The proposed method can recover from occlusion and tracking loss. Comparative experiments are presented to demonstrate the effectiveness and robustness of the proposed method.

1. Introduction

The purposes of this study are lips sensing are tracking and data acquisition of lips region. Recently, many of mobile devices, such as a cellular phone, PDA, and Micro Air Vehicle (MAV), have a camera. For these situations and ubiquitous computing, one of the most useful interfaces is the non-contact interface using images. However, there are few reports about computer vision using the MAV. This reason is that the acquired image is not stable. This difficulty is caused by instability of the MAV.

In this study, we focus to a lips region. The lips shape and color are the most important features of in the entire human race for anatomical reasons [1]. For these reasons, lips tracking and data acquisition is very important.

For mobile devices, some problems must be solved, which are listed below.

1. Complex situation (drastic change of whole scene)
2. Deformation of lips shape by speech
3. Acquisition of information of lips geometric changes
4. High speed and accuracy

The target image example is shown in Figure 1. This image is not only face region, has changes of whole scene by a camera motion, and includes lips deformation by speech (Figure 1(a) and (b)). The template image is only one closed mouth and prepared for each environment and person, because of personal use. The lips geometric information can be used for applications, such as correction of lips region for audio-visual speech recognition (Figure 1(c)). Moreover, the lips information which is acquired simultaneously with tracking for the real-time processing and simplicity.

A main technique of our method is Evolutionary Video Processing. In this method has a trade-off between accuracy and a processing time. In this paper, to solve this problem, we proposed Evolutionary Video Processing with automatic Search Domain Control (SD-Control). In comparative experiments, the effectiveness of the proposed method is verified. The proposed method improves the robustness. Moreover, the acquired data by the proposed method is useful for many applications, such as interface of MAV.

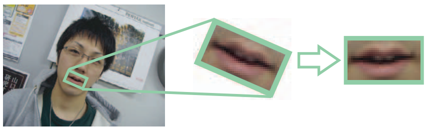

Figure 1. Basic concept: a) target image; b) results of detection, tracking, and sensing; c) application (correction of the lips).

2. Related Work

Some rule-based approaches [2] to measure lips movement have been proposed. These approaches cannot adapt to considerable geometric changes in every frame. In the model-based approach, Genetic Snakes [3] have been proposed. These approaches have some constraints, such as a helmet with a camera, because the initial setting of problems is needed and the number of nodes and parameters should be skillfully determined. Furthermore, real-time face and lips tracking system for facial expression recognition is reported [4]. These methods using whole face are difficult to be applied to our purpose. The reason being that the lips information which is acquired simultaneously with tracking for the real-time processing. Recently, the application of a particle filter to object tracking has become popular. The condensation algorithm [5] is proposed, which use the particle filter, the spline representation for contours of objects, and the affine transformation group parameters as the state vector. However, some reports indicate that these methods use a specific model of the object [6], moreover, particle filter requires an accurate model initialization [7].

To overcome these problems, we use template matching with a genetic algorithm (GA) in the proposed method. The proposed method uses not a special initialization but random numbers. Usually, a GA is unsuitable for a real-time processing. Moreover, the GA has a trade off between the exploration accuracy and the processing time. To overcome this tradeoff, Akashi et al. [8] have proposed Evolutionary Video Processing with flexible SD-Control. However, in this method, timing to control the search domain is fixed. This should be decided automatically.

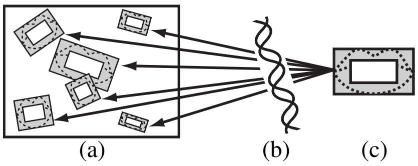

Figure 2. Concept of template matching with GA: a) target image with individuals; b) chromosome; c) template.

3. Lips Feature and Shape of Template

3.1 Lips Color and Shape of Template

The lips redness is a common feature 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 [1]. For this reasons, we use x component (redness) in the Yxy color space [9] as image data.

Usually, a template shape is square. This shape not suitable for lips shape deformation by speech. This is because, lips is deformed during speech, and has intense variations such as showing or not showing any teeth, as shown. To solve this problem, we focus our attention on invariance under constant shape deformations. Then, we find out that lips shapes of opened mouth during speech have the same topological properties [10]. Thus, we have used a template shape which has a ignored region to cope with the ever-changing lips region. This shape is called "square annulus". Here, w and h are defined as the source square template's width and height. w' and h' are width and height of inside ignored region. In this paper, in order to simplify the method, we fix the interior area dimensions; h' and w' are set to 80% and 50% of h and w respectively. These values are the empirical best values. Furthermore, there are the advantages of the "square annulus" that the ignored $w' \times h'$ region reduces the amount of calculation and makes the lips region extraction high speed.

4. Evolutionary Video Processing

In this part, the Evolutionary Video Processing is explained. A concept of template matching with GA is illustrated in Figure 2. A template is transformed by a chromosome of an individual (Figure 2(b)). Each individual is located on a target image, as shown in Figure 2(a), then these individuals are evaluated by a fitness function. These processes are explained below.

4.1 Structure of Chromosome and Fitness Function

A chromosome is a solution candidate to be optimized. Here, t_x and t_y are defined as coordinates after parallel translation, m_x and m_y are scaling rates, and $angle$ is rotation angle of lips shape. The template is transformed by these parameters on the target image.

Transformed templates (individuals) are evaluated by a fitness function. At first an objective value (O) is calculated, which represents a distance (pixel difference) between a template and a target image. The fitness function is shown as follow, $fitness = \max\{W_k, W_{k-1}, \ldots, W_{k-n}\} - O$, where W is the worst objective value, k is a current index of generation. The first term of is the worst objective value for last $n+1$ generations. This technique is called "scaling window" [11], and used for controlling selection pressure of GA. We use the window size $n = 5$ in the experiment which is decided from exploratory experiments.

4.2 Evolutionary Video Processing

Generally, in order to detect a moving object, an inter-frame difference picture is used as the information between video frames. However, it is difficult to use this in our method, because a camera moves intensively. Therefore, we use genetic information as a relation between video frames. In other words, without making new population, lips detection for a next frame proceeds with the population used in last frame. It is unthinkable that prodigious changes come out with geometric parameters, such as location, scaling, and rotation angle, in real-time processing. Therefore, this method can be expected to reduce the processing time and increase the accuracy.

Flow charts of the Evolutionary Video Processing is shown in Figure 3(a). In the Evolutionary Video Processing, an initial population is generated and after that the process flows the "Generation alternation" (Figure 3(b)). In Figure 3(a), special attention should be paid to the initialization of the GA population only in the first time. In other words, without initialization, the evaluation result in a previous frame is used in the next frame. This part is very simple, whereas this is important and effective.

5. Automatic SD-Control for Video Processing

5.1 Trade-off problem in GA

GAs are probabilistic search technique which are well-suited to the exploration of large and complex search spaces. There is a trade-off between search accuracy and speed. From our experiences, when the size of population and the number of generations are decreased, GA individuals get stuck at local optima as in Figure 4. This is because the GA is a global optimization algorithm, and not good for a local optimization [12]. As a search efficiency improvement, we can use a technique, in which after the domain including the optimal solution is specified, the neighborhood of that is searched in detail. However, this is a risky method. This is because the domain where the optimal solution is included clearly, cannot be specified. In contrast, in our many past experiments, we found that a part of a face is extracted as lips region, in case of the failure (Figure 4). This reason is that we use x component (redness) in the Yxy color space [9]. We hope that the Search Domain (SD) Control can escape from local optima and achieve a local optimization with a small population.

5.2 Automatic SD-Control

The search domain is controlled depending on both an elite individual and a condition of the evolution. The elite individual can be found out by comparison of the objective value of all individuals. The search domain center is set to a coordi-

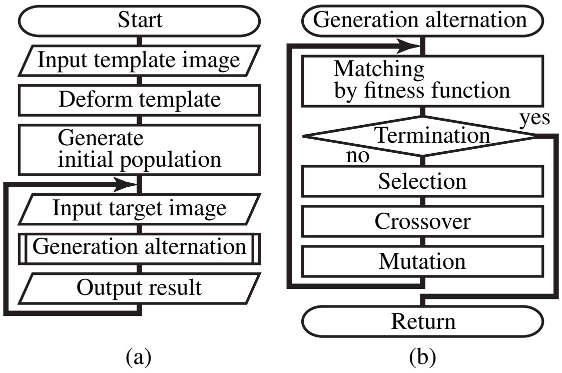

Figure 3. Flow charts of Evolutionary Video Processing: a) main process; b) GA process.

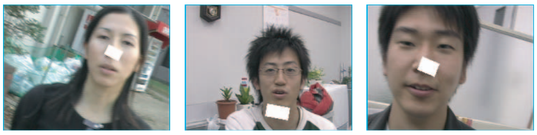

Figure 4. Examples of the local optimum.

nate (t_x and t_y) of the elite individual, whenever a new elite individual appears.

Next, the size of the search domain is decided by the condition of the evolution. The search domain is renewed as follows:

$$\begin{bmatrix} width^* \\ height^* \end{bmatrix} = \alpha \begin{bmatrix} width \\ height \end{bmatrix}. \qquad (1)$$

In Equation (1), $width$ and $height$ are the target image's width and height, and transformed to $width^*$ and $height^*$ respectively by α which is a scale factor, $\alpha = \{1.0, L, S\}$. At first, this α is set to 1.0, and then α is changed alternately L and S. The detection in the early stage of the GA is performed on the all target image, therefore this α is 1.0, at first. $\alpha = L$ and $\alpha = S$ mean that the GA searches on large area and small area respectively. By this mechanism, the search domain is incremented and decremented by the condition of the evolution. Therefore the search domain is controlled automatically. The search domain is incremented and decremented to achieve an efficient search by a small population and escape from local optima.

Timings to change the scale factor α is also important. The proposed method determines the timing depending on the condition of the GA evolution. The condition of the evolution means the evolution proceeds or reaches a plateau. The scale factor α repeats change L and S every time the evolution reaches a plateau (or convergence). The number of generations which the same elite individual continues, determines whether the evolution reaches the plateau or not. These thresholds are defined as follows; the threshold of changing α to L is T_L, and S is T_S.

These S, L, T_L, and T_S are determined empirically in this paper. The values of T_L and T_S which are used in experiments, are shown in Section 6. For improvement of decision of these is one of the our future work.

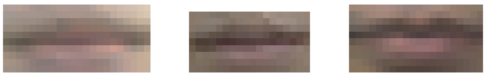

Figure 5. Examples of template images (closed mouth).

In order to apply the automatic SD-Control to Evolutionary video processing, the proposed method inherits not only the genetic information but also the search domain.

6. Experiments

6.1 Input Images

At first, we input one template image, after that, target video frames are read sequentially. The number of subject is 4, and 4 scenes are taken (two scenes for each indoor and outdoor situation). Therefore, 16 video sequences are used in experiments. The template is closed mouth and prepared 16 images for every scene. Examples of the template image are illustrated in Figure 5. The average size of the template image is 20.75×10.81 pixels.

A frame size of the target video sequence size is 240×180 as shown in Figure 1(a) and 4. These are trimmed from a video sequence, which is taken by a digital video camera. Some red color objects are included in the background, such as people in a poster, some flowers, and so on. The subjects repeat pronunciation of the Japanese vowels. Moreover, on the assumption that the camera moves and joggles by free hand, the shaking of scene is caused artificially by hand. For the simulations, we use five seconds video sequence (150 frames) with 30 frames per second.

6.2 Parameters

Parameters and settings of the both GAs are as follows. The population size is small 10, the crossover and mutation probability is 0.7 and 0.2. The elite preserving strategy is used, and the type of crossover is uniform crossover.

The relationship between the search speed and accuracy should be investigated for practical use. Therefore, in addition to comparison with method A and B, we compare the accuracy between the long alternation of generations and short ones. The difference of the random seed must have no effect on the comparison. Therefore one random seed is used for both methods A and B to compare under the same conditions. In this part, 16 target video sequences are prepared. The trial is 5 times for each video sequence. Therefore 80 initial random seeds are prepared. Parameters of the automatic SD-Control are $L = 0.5$, $S = 0.25$, and $T_L = T_S = 10$.

6.3 Accuracy and Processing Time

In this part, the accuracy and processing time is evaluated. A computer, with a Pentium4 3.0GHz CPU, is used.

The accuracy and processing time results of the method A and B are shown in Table 1. The results of lips region tracking is judged to be good or not good by visual observations. If the tracked area is covered the whole lips region, the tracking is judged to be good. At first row in Table 1, results of the method A, which the GA is terminated at 200 generations are

Table 1. Tracking accuracy and processing time.

method	GA	accuracy (%)					time
	term.	sub 1	sub 2	sub 3	sub 4	avg	(msec)
A	200 th	93.5	92.1	93.9	98.4	94.5	32.8
A	50 th	77.6	82.8	84.4	94.8	84.9	8.4
B	50 th	58.4	72.6	50.5	83.7	66.3	

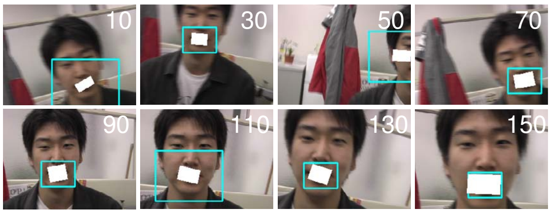

Figure 6. Example of results: disappearance and zooming.

shown. The accuracy is very high 94.5%, and the processing time is 32.8 milliseconds.

The method A is compared with the method B to evaluate the effectiveness. Second and third row in the Table 1 show accuracies of the method A and the method B. In these experiments, the GA is terminated at 50 generations. The method B is very low, 66.3%, and the method A is 84.9%. This method A accuracy is lower than the first row in Table 1. Comparing the method A and the method B with 50th generations in Table 1, the method A is better than the method B for all video sequence. The processing time of one frame is a quarter of Table 1, about 8.4 milliseconds. It is clear that performance of the proposed method is better than the conventional method. These results show that the proposed method can efficiently explore with a smaller computational effort. This means the Evolutionary Video Processing is downsized.

6.4 Robustness and Data Acquisition

In this part, the robustness of the proposed method is evaluated. Figure 6 illustrate results of disappearance from a frame and zooming. The number on the top-right corner in these figures is the number of the frame.

Figure 6 shows examples of the results in a situation, which a lips region disappears from a frame and a camera moves closer to a subject. These results are selected every 20 frames from a resulting video sequence. In the 50th frame, the lips region disappears from the target frame, however the tracking is recovered. Moreover, in the latter half of the sequence, the camera gradually moves closer to a talking subject. Eventually, the target video frame reaches the 150th frame whose true solutions are manually measured as shown in Table 2. The scaling rate of width and height are 3.09 and 3.76. In previous sections, the scaling rates are coded in 8-bit by range of $[0.8, 3.0]$. This range cannot apply in this situation. Therefore the range is changed as $[0.8, 5.0]$. Table 2 shows numerical results by GAs of the method A and manual. Both solutions which are found by the method A and manual are nearly equal. This means that the method A can support the significant geometric changes by using a large range in accordance with the intended use. For some applications devices described in Section 1, such as audio-visual speech recognition, speaker identification system, robot perception, interface of personal mobile, the lips can be corrected using these acquired lips data in Table 2.

Table 2. Numerical results in the 150th frame in Figure 6.

method	t_x	t_y	m_x	m_y	$angle$ [deg]
manual	127	138	2.99	3.73	3.2
GAs	125	136	3.09	3.76	2.1

7. Summary

In this paper, high speed lips tracking and data acquisition of a talking person in natural scenes were presented. Our approach is based on the Evolutionary Video Processing. This method has a trade-off between accuracy and a processing time. To solve this problem, we proposed and implemented the automatic SD-Control to the Evolutionary Video Processing. In our simulations, the effectiveness and robustness of the proposed method are verified by a comparison experiment. It is demonstrated that the lips region detection and tracking at high speed and with high accuracy is possible, simultaneously with acquisition of its numerical geometric change information. Our future work is application the proposed method to the audio-visual speech recognition or the interface of mobile devices.

References

[1] L. L. Otis, D. Piao, C. W. Gibson and Q. Zhu: “Quantifying labial blood flow using optical doppler tomography”, Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology, **98**, 2, pp. 189–194 (2004).

[2] C.-C. Chiang, W.-K. Tai, M.-T. Yang, Y.-T. Huang and C.-J. Huang: “A novel method for detecting lips, eyes and faces in real time”, Real-Time Imaging, **9**, 4, pp. 277–287 (2003).

[3] L. Ballerini: “Genetic snakes for medical images segmentation”, Proceedings of SPIE Mathematical Modeling and Estimation Techniques in Computer Vision, Vol. 3457, USA, pp. 284–295 (1998).

[4] N. Oliver, A. Pentland and F. Bérard: “Lafter: a real-time face and lips tracker with facial expression recognition”, Pattern Recognition, **33**, 8, pp. 1263–1403 (2000).

[5] M. Isard and A. Blake: “Condensation – conditional density propagation for visual tracking”, nternational Journal of Computer Vision, **29**, 1, pp. 5–28 (1998).

[6] D. Serby, E. K. Meier and L. V. Gool: “Probabilistic object tracking using multiple features”, Proceedings of the 17th IEEE International Conference on Pattern Recognition (ICPR 2004), Vol. 2, UK, pp. 184–187 (2004).

[7] M. W. Lee, I. Cohen and S. K. Jung: “Particle filter with analytical inference for human body tracking”, Proceedings of IEEE Workshop on Motion and Video Computing (MOTION2002), USA, pp. 159–165 (2002).

[8] T. Akashi, M. Fukumi and N. Akamatsu: “Real-time genetic lips region detection and tracking in natural video scenes”, Proceedings of the 2004 IEEE Conference on Cybernetics and Intelligent Systems (CIS 2004), Singapore, pp. 682–687 (2004).

[9] K. N. Plataniotis and A. N. Venetsanopoulos: “Color Image Processing and Applications”, Springer-Verlag, Germany (2000).

[10] F. Klein: “Erlangen program”, Inaugural address at the University of Erlangen (1872).

[11] D. Whitley: “The GENITOR algorithm and selection pressure: Why rank-based allocation of reproductive trials is best”, Proceedings of the Third International Conference on Genetic Algorithms (ICGA’89), USA, pp. 116–121 (1989).

[12] D. E. Goldberg: “Genetic Algorithms in search optimization & Machine lerningn”, Addison-Wesley Publishing Campany, Inc., Boston, MA, USA (1989).