SICE-ICASE International Joint Conference 2006
Oct. 18-21, 2006 in Bexco, Busan, Korea

Vowel Recognition System by Lip-Reading Method Using Active Contour Models and its Hardware Realization

Satoru NAKAMURA[1] , Takao KAWAMURA[2] and Kazunori SUGAHARA[2]

[1]Graduate School of Engineering, Tottori University, Tottori, Japan
(E-mail: snakamur@ike.tottori-u.ac.jp)
[2]Department of Engineering, Tottori University, Tottori, Japan
(E-mail: {kawamura, sugahara}@ike.tottori-u.ac.jp)

Abstract: Even in noisy environments such as in machinery factories or in crushes, we sometimes wish to control equipments by using human voices. Although word recognition methods have been successfully developed, noises in environments cause serious recognition problems. In such cases, lip shape movements are expected to be useful as the supporting data to improve the performance of recognitions. In this paper, a method to extract outer and inner lip shapes from input face images by using Active Contour Models is proposed. The extracted lip shape data are utilized to control equipments by lip-readings in noisy circumstances. Normalization method to reduce the effect of the lip size change depending on the distance from a camera to human faces, and to improve the recognition rate is also mentioned. A three-hierarchy Neural Network is used for the recognition method. The proposed method is implemented as hardware circuits in a FPGA chip to process NTSC video signals in real-time. Experimental results of vowel recognition are shown to confirm effectiveness of the proposed method.

Keywords: Active Contour Models, Vowel recognition, Hardware realization, FPGA.

1. INTRODUCTION

Word recognition methods have been developed, so that even unspecified voices uttered by unspecified person can be recognized in high recognition rate. However, in noisy environments, noises cause serious recognition problems and recognition rates become low as a result. But word recognition is practically expected to be utilized in noisy environments. Considering these points, supplemental use of visual information of lip shape movements is expected to improve the performances. In order to use lip shape movements for word recognition, a fast and accurate extraction method of lip shapes from face images is required.

Area extraction technique is able to applied to lip shape extractions. Area extraction is one application of area extraction techniques and is developed by several methods such as spatial filtering methods. Active contour model, originally proposed by Kass et al., is one of such area extraction techniques. Kass's Active Contour Model (Snakes) solves image energy minimization problems. Various features of images, such as color of pixels or sharpness of specified areas, can be considered as image energies in Snakes, and flexible area extraction systems are easily constructed.

Area extraction systems can be realized flexibly by using Snakes, however, Snakes require large amount of calculations and long computational time to solve energy minimization problems. Large amount of calculations and long computational time make it impossible to implement area extraction functions based on Snakes in stand-alone system. Considering these points, Hashimoto et al. proposed a new type of Active Contour Model named as Sampled-ACM to decrease computational costs[1]. In this model, area extraction problems are assumed as force balancing problems of sampling points on the closed curves. By using the Sampled-ACM, fast area extraction is available because they just calculate sum of forces which works on each contour points.

A new force called a vibration factor is introduced by Sugahara et al. for improving accuracy against noises in images and they have tried to realize Sampled-ACM with vibration factor as hardware circuits in FPGA (Field Programmable Gate Array)[2]. The Sampled-ACM realized as hardware circuits makes easy implementation of area extraction function in the stand-alone systems, however, it extracts only one area in the images.

In this paper, the Sampled-ACM is used to extract outer and inner lip shapes. Moreover, to reduce the influence on the recognition rate according to the distance from a camera to human faces, normalization method of the extracted lip shapes is discussed. The normalized lip shape data are recognized by using a three-hierarchy Neural Network. Finally, the proposed method is implemented as hardware circuits in a FPGA chip for easy development of stand-alone system and for processing NTSC video signals in real-time. The experimental results of vowel recognition are shown to confirm effectiveness of the proposed system.

2. SYSTEM CONFIGURATION

The system configuration is shown in Fig.1. Face image at the moment of utterance is taken from the NTSC video camera as an input of the system, and the image is digitized by NTSC video decoder and is stored in the memories on the FPGA board. Active Contour Model is applied to the image stored in the memories immediately after storing of one frame of image, and the lip shape is extracted. As a result, the coordinate data of the operation points of ACM are obtained and they express the lip shape. After that, they are normalized to reduce the effects of lip shape size and to improve the recognition

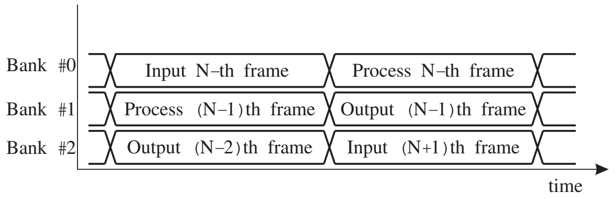

Fig. 2 Three bank memory

rate. The normalized data are sent to the vowel recognition circuit, and the recognition result is output. Moreover, the converged ACM results are combined with original input image, and they are output to the TV monitor through NTSC encoder to confirm that the extraction is succeeded.

In this system, the memories are constructed as three banks and these banks can be read/written by FPGA simultaneously. By using three bank memories, inputting, processing and outputting data are accomplished at the same time as shown in Fig.2. Each frame of images from the NTSC video camera are decoded by the decoder LSI and the obtained digital data are stored in one bank of memory on the FPGA board. The Sampled-ACM is applied to the frame stored in a bank of memory. Processing results stored in a bank memory are output through the NTSC encoder LSI.

3. LIP SHAPE EXTRACTION BY USING ACTIVE CONTOUR MODELS

Active Contour Models are energy minimizing closed curves controlled by external constraint forces. To improve the recognition rate, not only the outer lip shape but also the inner lip shape are used as input data for the vowel recognition. The initial Sampled-ACM is set to locate all around the image and color information of target area are set in the ACM before the application of it.

After the binarization of the input image according to the color information, first the outer lip shape is extracted by using the Sampled-ACM and the inner lip shape can be obtained by continuously applying the Sampled-ACM to reversed input image.

4. NORMALIZATION

After the lip shape extraction, the coordinate data of the obtained operation point of Sampled-ACM should be normalized. Since the raw coordinate data of the converged Sampled-ACM have different values according with the lip shape size and with converged conditions, the recognition process cannot obtain successful results without data normalization. The normalization procedures adopted in this paper are described as follows.

1. Obtain two contour points which have the maximum/minimum values of x coordinate data.
2. Obtain the center point of the converged lip shape. The center point is culculated by dividing the line that links above two points in half.
3. Draw straight lines at every 30 degrees from the center point.
4. Connect the adjacent points with straight lines.
5. Calculate the intersections of 3 and 4, and obtained values are coordinate values of normalized lip shape.

The normalization process is summarized in Fig.3.

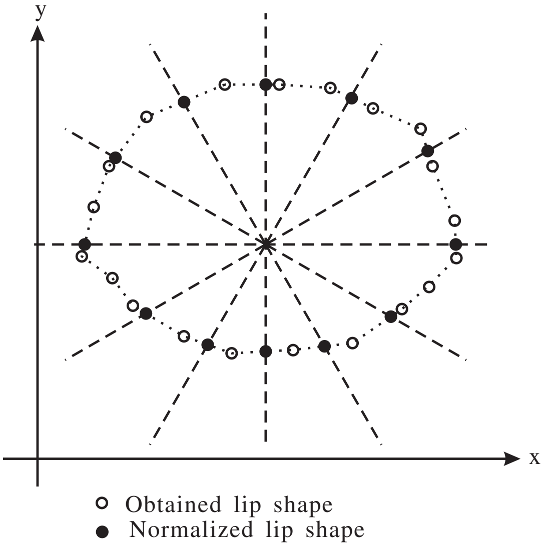

Fig. 3 Normalization

5. VOWEL RECOGNITION METHOD

In this paper, a three-hierarchy Neural Network is used for the vowel recognitions. And the back propagation algorithm is used for training of the Neural Network. The three-hierarchy Neural Network is shown in Fig.4. The normalized lip shape data are input into an input layer, and the result of vowel recognition is given by the unit number which has the maximum output value among the units in the output layer.

6. HARDWARE REALIZATION OF VOWEL RECOGNITION SYSTEM

Hardware realization of the proposed system is accomplished on FPGA (Field Programmable Gate Array) chip and VHDL is used to describe the circuits. FPGA used in the proposed system is Cyclone EP1C20F400C7 fabricated by the ALTERA Co. and its logic Element number is 20,060. These specifications of FPGA are summarised in Table 1.

Table 1 FPGA specifications

Manufactured	ALTERA Co.Ltd.
Model number	EP1C20F400C7
Number of logic elements	20,060
Max. I/O port number	301
Package	400-Pin Fine Line BGA
Package size	21 × 21 [mm]
Frequency	48[MHz]

The equipments used in this system are shown in Fig.5. Quartus II software is utilized as the developing

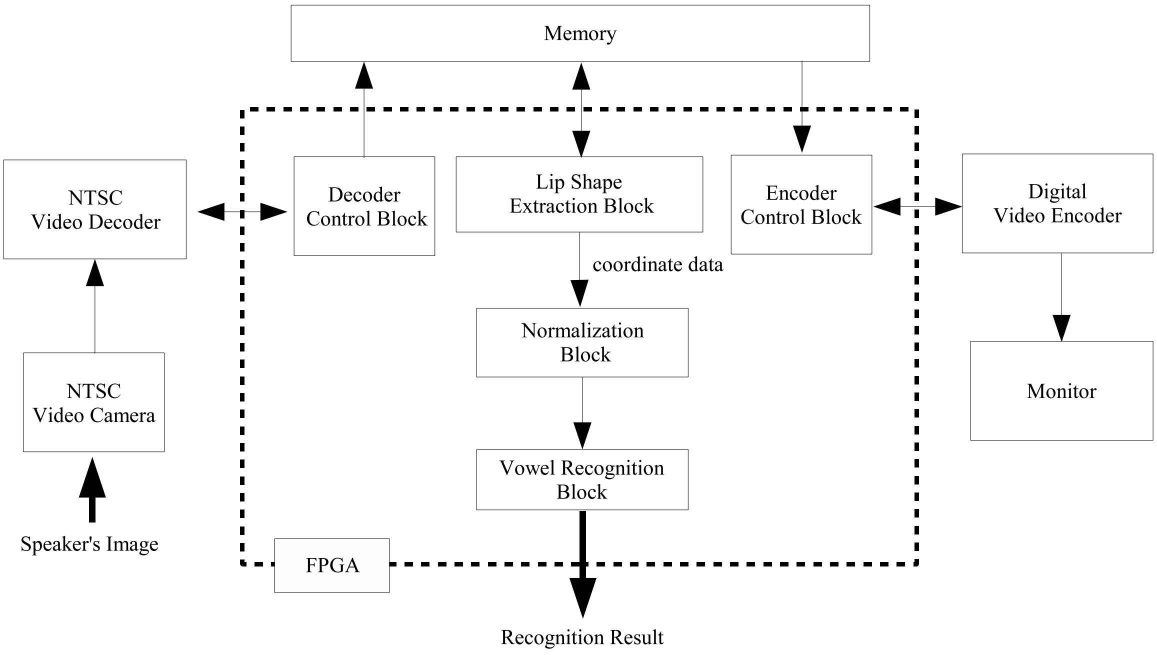

Fig. 1 System configuration

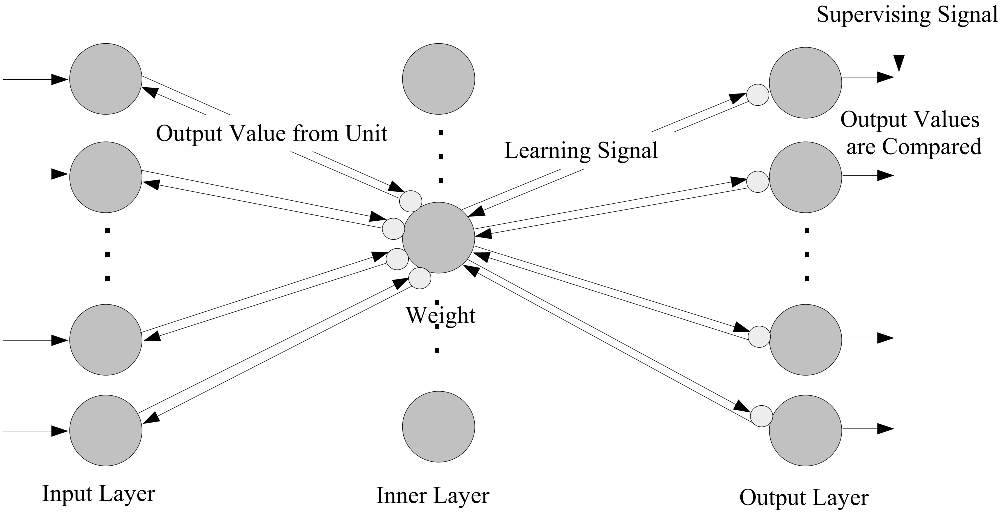

Fig. 4 Three-hierarchy Neural Network

tool of FPGA circuits. Adders and multipliers for the calculation of the floating point numbers are designed by using the Mega Wizard functions of Quartus II.

7. EXPERIMENTS

Original FPGA board is designed for the proposed system. On the FPGA board, three bank memories, video encoder and video decoder are equipped. The three bank memories can be accessed from three circuits simultaneously and provide fast parallel signal processing.

Figure 6 shows the example of extracted lip shape. As shown in the figure, outer and inner lip shapes are extracted successfully. In this figure, black and white dots show the extracted outer and inner lip shapes, respectively.

Vowel recognition experiments are examined by using the developed vowel recognition system. 50 human face images of uttering five Japanese vowels and 50 images without uttering are prepared. The total number of images used in the experiments are 300. The recognition results obtained by the experiments are summarized in the Table 2. 72.3% of the total recognition rate is achieved as shown in the Table 2.

Required number of clocks and processing time at 48MHz system clock for each processing of 1 frame are summarized in the Table 3. As shown in this table, total processing time for 1 frame is approximately equal to 7[msec] and is enough to process video signals in real time.

8. CONCLUSION

In this paper, the method for vowel recognition based on the lip shape movements are described. The recognition of the lip shape movements are referred as the

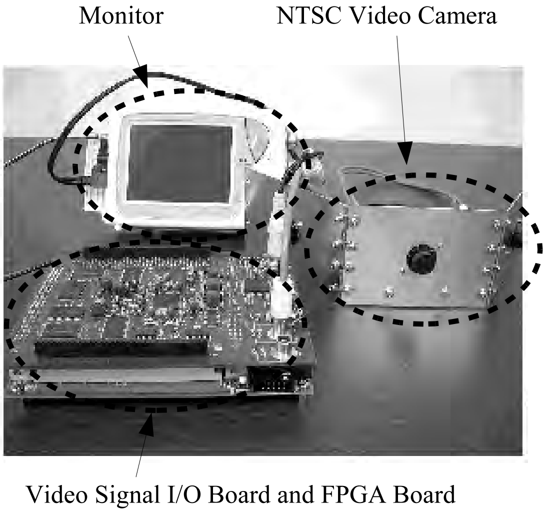

Fig. 5 System configuration

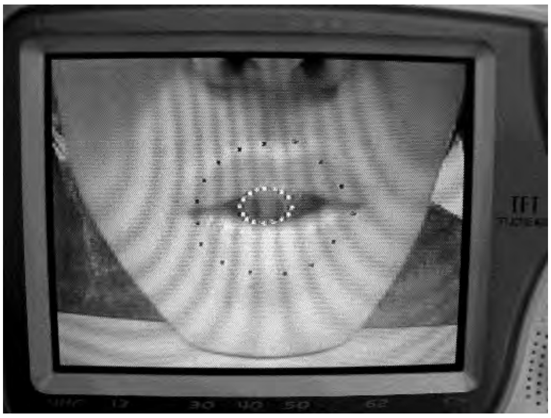

Fig. 6 An example of the extracted lip shapes

lip reading and is achieved by using the Sampled-ACM. Hardware realization of the proposed system on FPGA chip is also shown. The human face images at the moment of uttering are used as input data and are recognized. The normalization method to reduce the influence according to the the lip shape size and to improve the recognition rate is also mentioned. The three-hierarchy Neural Network is used for the vowel recognition.

REFERENCES

[1] M.Hashimoto, H.Kinoshita and Y.Sakai, " An Object Extraction Method Using Sampled Active Contour Model," IEICE Trans. D-Ⅱ, Vol.J77-D-Ⅱ, No.11, pp.2171-2178, 1994.

[2] K.Sugahara, T.Shinchi and R.Konishi, "Active Contour Model with Vibration Factor," IEICE Trans. D-Ⅱ, Vol.J80-D-Ⅱ, No.12, pp.3232-3235, 1997.

[3] Y.Sasaki,T.Kawamura and K.Sugahara,"Hardware Realization of Lip Shape Extractions by Using Active Contour Model and its Application for Word Recognition," The 6th IEEE Hiroshima Student Symposium ,pp.171-174, 2004.

Table 2 Vowel Recognition Result

Vowel	Success	Failure	Recognition rate (%)
a	40	10	80.0
i	25	25	50.0
u	44	6	88.0
e	36	14	72.0
o	34	16	68.0
not speech	38	12	76.0
Total	217	83	72.3

Table 3 Required number of clocks and time

Processing	Number of clocks	Processing time [msec]
Lip shape extraction	267,000	5.56
Data normalization	11,000	0.23
Vowel recognition	51,320	1.07
Total	329,320	6.86