Hardware Realization of Panoramic Camera with Direction of Speaker Estimation and a Panoramic Image Generation Function

TORU UESUGI
Graduate School of Engineering
Tottori University
Koyama-Minami, Tottori
JAPAN
s022011@ike.tottori-u.ac.jp

TAKAO KAWAMURA
Faculty of Engineering
Tottori University
Koyama-Minami, Tottori
JAPAN
kawamura@ike.tottori-u.ac.jp

TADAAKI SHIMIZU
Faculty of Engineering
Tottori University
Koyama-Minami, Tottori
JAPAN
tadaaki@ike.tottori-u.ac.jp

KAZUNORI SUGAHARA
Faculty of Engineering
Tottori University
Koyama-Minami, Tottori
JAPAN
sugahara@ike.tottori-u.ac.jp

Abstract: In most of the current teleconferencing systems, a conference room is taken by one fixed video camera. For that reason, face images of speakers become small. To take both full view images of a conference room and face images of speakers, it is required to develop a multi camera system with human operations. In this paper, we develop a new video camera for a teleconferencing in above mentioned situations. The proposed video camera is constructed with three NTSC video cameras, and it outputs the images composed with panoramic images generated by images from these video cameras and the facial images of speakers. In this paper, an Active Contour Model is adopted for panoramic images, directions of speakers are estimated by voice and hardware realization of the system is mentioned.

Key–Words: Direction of Speaker Estimation, An Active Contour Model, Splitting characteristics, Panoramic image, TV conference, FPGA

1 Introduction

The advancement of communication speed in digital networks makes it possible to exchange a large amount of information between distant places. Teleconferencing is one typical example of the practical applications in such fields. In most of the current teleconferencing systems, a full view image of a conference room and voice of speakers are bidirectionally exchanged between distant places. However, in teleconferencing situations, face images of speakers are also requested to perceive facial expression. To take both full view images of a conference room and face images of speakers, it is required to develop a multi camera system with human operations. Therefore we develop a video camera for teleconferencing to output the images which are composed with panoramic images generated by images from three NTSC video cameras and the facial images of speakers. In order to constitute such a device, some important techniques are required such as a technique to extract multiple speaker areas from a panoramic image, to estimate speakers directions. For a method to extract a speaker area, we apply an Active Contour Model[1]. However, we need to extract multiple speaker area from one image because more than one person speak at a time by letting a closed curve of a conventional an Active Contour Model divides.

In this paper, an Active Contour Model is applied to the camera image chosen as results of the direction of speaker estimation processing by voice, multiple speaker areas are extracted. Processing that displays combination of enlarged speaker's images and panoramic images. Hardware realization of above system in an FPGA chip is also investigated.

2 Hardware configuration of the proposed system

The proposed system is constructed with three NTSC video cameras and an FPGA board constructed with NTSC video encoder/decoder LSIs and memories as shown in Fig.1.

The proposed panoramic image generation function is realized on an FPGA chip as hardware circuits for easy constructions of embedded real-time systems. An FPGA is LSIs that have reconfigurable inner circuits. The operation clock applied to the FPGA is 48MHz. The Logic Element number is 20,060. These specifications are summarised in Table 1.

In this system, the memories are constructed as three banks and these banks can be read/written by an 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 cameras is de-

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 FineLine BGA
Package size	21 × 21 [mm]
Frequency	48[MHz]

coded by the decoder LSI and the obtained digital data are stored in one bank of memory on the FPGA board. The proposed panoramic image generation function is applied to the frame stored in a bank of memory.

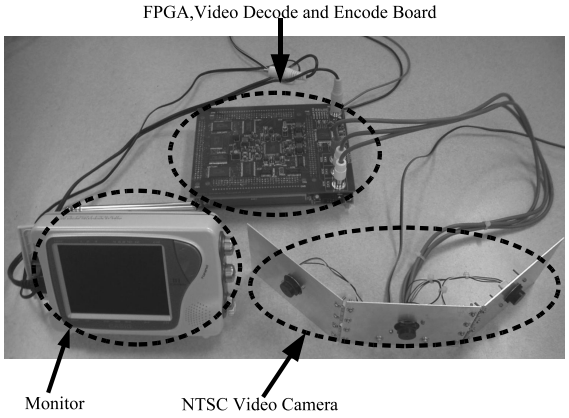


Fig. 1: System configuration.

3 Generation of panoramic images

The input images from the three video cameras are compliant with the NTSC specifications, that is, they are 640×480 [pixels] size and 30 frames per second. The output image also should be compliant with the specifications, i.e. it should have the same size and the same frame rate.

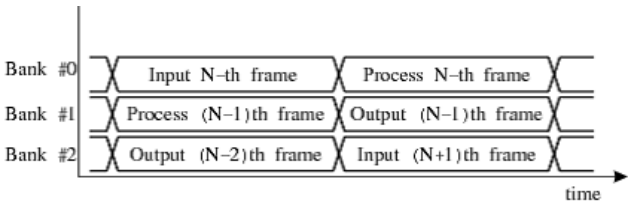


Fig. 2: Three bank memory.

As shown in the figure, the lower half area of the output image is arranged for the panoramic image and the extracted speaker face image is enlarged and is represented on the upper half of it. The panoramic image part in the output image is 640 × 180 [pixels] size and is generated with $3/8$ reduced 3 input images with two overlapped area in 40 pixel width as shown in Fig.3.

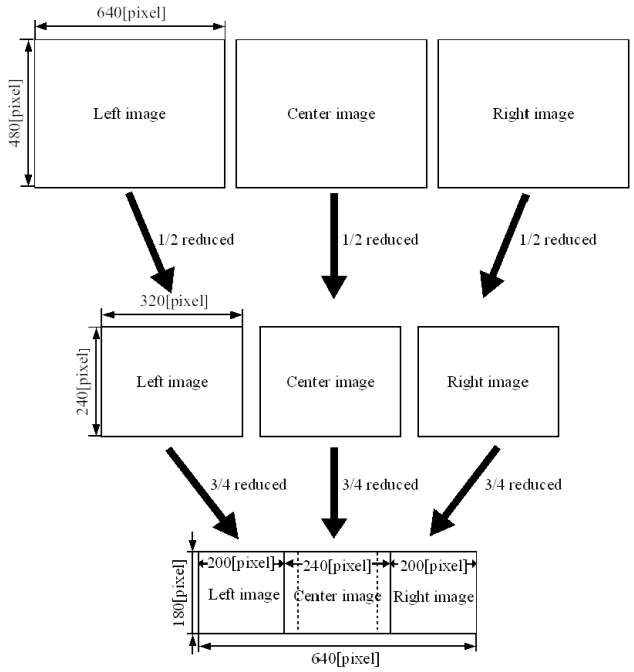


Fig. 3: Panoramic image generation before correction.

4 Multiple area extraction technique

4.1 Four forces work on contour points of an active contour model.

An Active Contour Model is a virtual closed curve with some contour points to extract a specified area from images. However, when it comes to extract multiple face areas from one image, this method is not effective. Then, multiple particular fields are extracted from the image by splitting the closed curve. Fig.4 shows the appearance where a closed curve is shrunk by four forces which work on contour points of a closed curve.

The pressure F_p affects every contour point, and F_p points in the direction to bisect the angle made of the target point and its adjacent points. The magnitude of F_p is represented as K_p which is the pressure constant. The two attractions F_{a1} and F_{a2} also affect

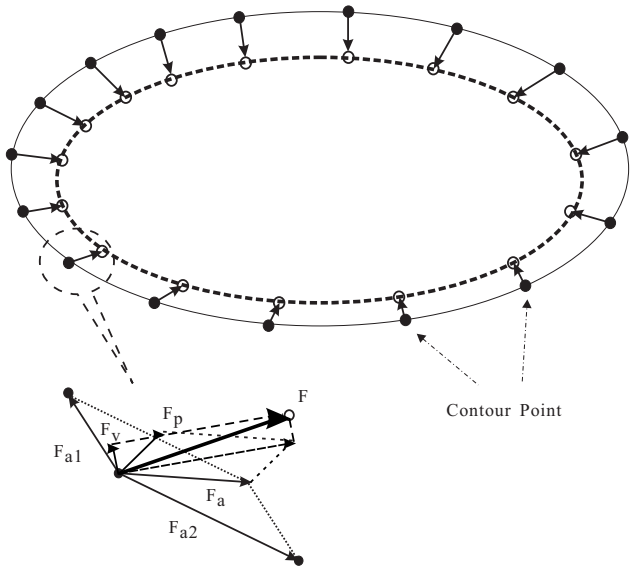


Fig. 4: Retraction of closed curve using an active contour model.

every contour point and their magnitude are proportional to the distance between the target point and its each adjacent contour point. The vibration factor F_v has constant magnitude and it works perpendicularly to the sum of F_p and $F_a = F_{a1} + F_{a2}$.The direction of the vibration factor reverses each turn of convergence. When a contour point hits the edge of a specified area in an input image, the repulsion F_r works to cancel the vertical component of $F = F_p + F_a + F_v$ and the contour point stays at the edge of the area as shown in Fig.5.

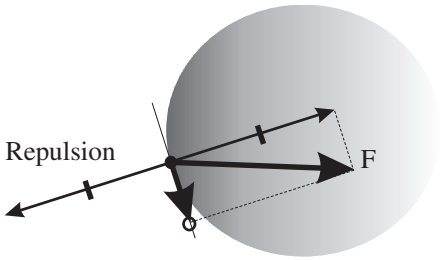
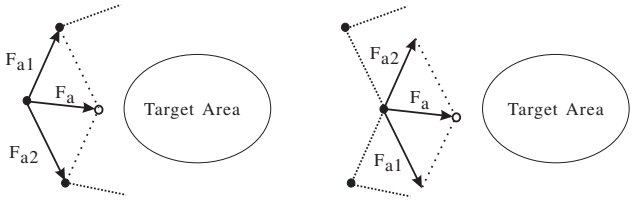


Fig. 5: Repulsion.

4.2 Attraction

When the closed curve becomes convex internally as shown in Fig.6, attraction reverses the direction of power to add the contour point, and does the operation that becomes power to which the power drawn to the adjoined contour point is pushed out. As a result, the retraction of the closed curve by pressure and attraction becomes possible, and the extraction of bumpy target area becomes possible. Furthermore, it becomes possible to keep the balance of the closed curve at the retraction only of attraction. Pressure value K_p is assumed so little that the chance to go into the position that becomes convex from two contour points that the contour point to which it pays attention adjoins internally is made. Thus, an Active Contour Model can be comparatively expressed by a simple mechanism by the use of this attraction.



(a) conventional attraction (b) attraction in this paper

Fig. 6: Attraction.

4.3 Split active contour model

Sole area in images is able to be extracted by means of the conventional Active Contour Model. However, when targets area are multiple in images, these areas are not able to be extracted correctly by using the conventional model. Here, split Active Contour Model show the appearance in which multiple areas are extracted.

Fig.7(a) shows the initial Active Contour Model. Fig.7(b) shows the Active Contour Model before splitting behaviour and the same result can be obtained by using conventional Active Contour Model. Measuring distances between contour points, when they become shorter than certain distance given as threshold value h, an Active Contour Model starts splitting behaviour as shown in Fig.7(c), (d) and (e). The same process continues until getting final results as shown in Fig.7(f).

5 Direction of speaker estimation technique

There are some techniques to estimate directions of speakers, but three directions are estimated equivalently to the number of cameras in this system. Therefore, high angle resolving power is not required but a simple following techniques for processing on an FPGA is sufficient.

First of all, ideal sine wave received by microphone is shown in eq.(1). Here, Fs is the sampling

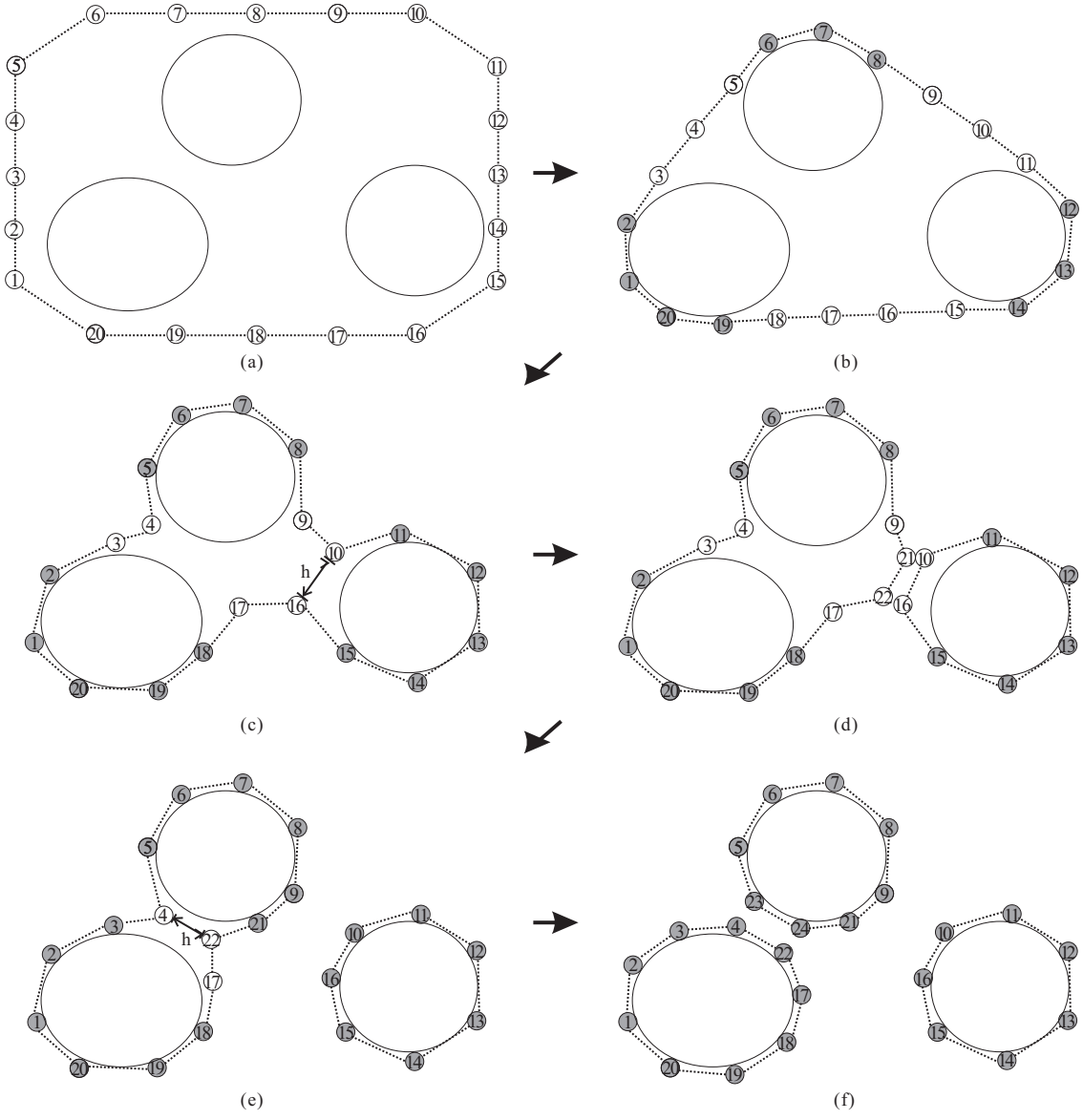


Fig. 7: Split ability of Active Contour Model (a) Initial state, (b) Before splitting behaviour, (c) Some contour points locate between two objects, (d) First splitting behaviour, (e) Second splitting behaviour, (f) Final results.

frequency.

$$\sin\Big(2\pi f\frac{t}{Fs}\Big) \tag{1}$$

The phase difference $\Delta\phi$ that is by delaying only at time τ in eq.(1).

$$\Delta\phi = \frac{2\pi f\tau}{Fs} \tag{2}$$

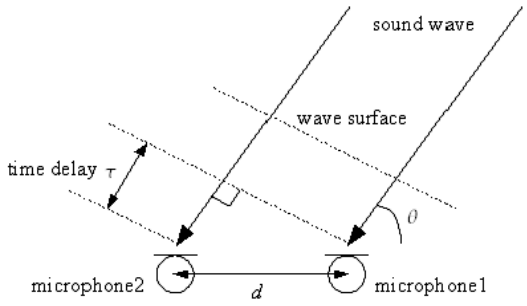


Fig. 8: Direction of estimation model.

When sound waves are received by two microphones spatially arranged as shown in Fig.8, time lag occurs between two data received by microphones. These sound waves are received by two microphones located at distance of d.

The sound wave that comes from the direction of θ in Fig.8 is received in microphone1. Next, the sound wave reaches microphone2 with the delay time τ shown in Fig.8. This relation is expressed by eq.(3). In this equation, constant c is the speed of sound waves.

$$\tau = \frac{d\cdot\cos(\theta\times\frac{\pi}{180})\cdot Fs}{c} \tag{3}$$

When eq.(2) and (3) are solved about direction of sound source θ, it becomes the following.

$$\theta = \cos^{-1}\Big(\frac{\Delta\phi c}{2\pi f d}\Big)\times\frac{180}{\pi} \tag{4}$$

The direction of speaker can be estimated by the calculation result of the eq.(4).

6 Output image generation processing

Concerning speaker area extraction and output image generation processing from panoramic image, the following processes are accomplished.

1. Images are taken with the 3 NTSC video cameras and the split Active Contour Model is applied to one of the three input images of 320×240[pixel].
2. The enlarged facial image is generated based on the speaker area extraction results.
3. The panoramic image is generated from three input images.
4. The output image is synthesized from the speaker image and the panoramic image.

Fig.9 shows the specification of the synthesized output image.

The flow of the processing of the expansion display is as follows. First of all, split Active Contour Model is applied to the input image. The mean value of x coordinates and y coordinates of contour point in 20 points that extracts the speaker area is calculated. Afterwards, the speaker image is expanded by centering on the coordinates point of the calculated mean value. However, the image of 213×80[pixel] of the input image that reduces to $3/8$ size 240×180[pixel] is taken out in consideration of the image size of the output image (640×480[pixel]), and 639×240[pixel] that is the three times is output to the monitor as a speaker image.

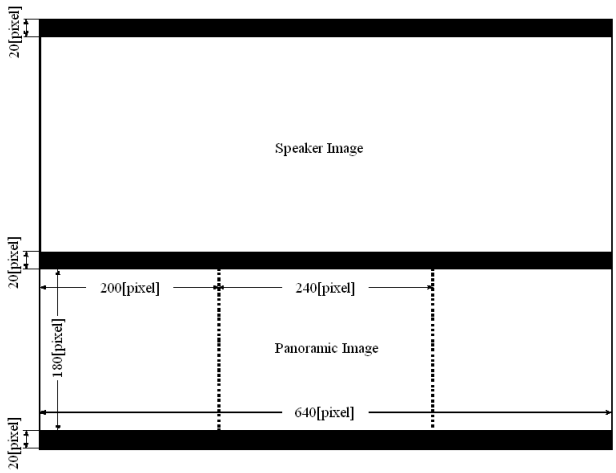


Fig. 9: Output image.

7 Experiment

7.1 Direction of estimation experiment

The result of the direction of estimation is shown in Table 2 to confirm the effectiveness of the proposed technique. The sampling frequency Fs is set

as 44.1[kHz] and the frequency of input sound wave is f=1[kHz]. The sound data from various angles are taken. Moreover, the input sound wave is assumed to be a plane wave at microphones. Two microphones are placed with distance d=6[cm]. The voices used are recorded in the conference room where the adjustment of a special sound characteristic is not given. The distance between the sound source and the microphone is set to 1[m].

Table 2: The result of the direction of estimation.

θ[deg]	τ[s]	experimental value[deg]
20	4.5×10^{-5}	14.8
35	9.1×10^{-5}	31.0
145	-9.1×10^{-5}	149.0
160	-3.4×10^{-5}	168.9

7.2 Image synthesis processing experiment

An experimental result of image synthesis processing is shown in Fig.10. Small square points in Fig.10 are shrunk contour points. Fig.10 shows split Active Contour Model can extract multiple areas, that is, two facial areas in this experiment. The developed circuit was able to be achieved by about 67% of the number of logic elements that was able to be used. Clocks required for processing of each frame are about 6×10^6. In other words, 123 [msec] is required for processing each frame.

Fig. 10: Experimental result of image synthesis processing.

8 Conclusion

In this paper, the hardware realization of the system that displays the panoramic image of the conference scenery and the speaker's facial image is developed video camera for teleconferencing. This system has extracted multiple speaker images with applying split Active Contour Model to the panoramic image. And the hardware realization of processing that estimated the direction of speaker is accomplished.

References:

[1] T. Miaki, T. Kawamura and K. Sugahara, Hardware Realization of Active Contour Models with new attraction for Multiple Area Extraction, *FIT Trans.* 2006, pp. 261–264.

[2] R. Yonemoto, T. Uesugi, T. Kawamura and K. Sugahara, Video Camera for Teleconferencing with a Panoramic Image Generation Function, *IEEJECT Trans.* Vol. ECT-06-119, 2006, pp. 31–35.

[3] Y. Li, R. Katsura, K. Sugahara and R. Konishi, Direction of Arrival Estimation of Multiple Voice Sources by Using Frequency and Phase Estimation Methods, *IEICE Trans.* A, Vol.J86-A, No.3, 2003, pp. 320–325.