Paper

Moving Pictures Segmentation and Adaptive Motion Compensation

Non-member Zhigang Chi (Himeji Institute of Technology)
Non-member Toshifumi Kimura (Himeji Institute of Technology)
Member Kenji Yamauchi (Himeji Institute of Technology)
Non-member Kennichi Hatakeyama (Himeji Institute of Technology)

In this paper, we propose a method to segment images into suitable number of contents by motion information, and merge the contents with the same or almost the same motion to one to decrease the overhead and increase the compression ratio. For content with the global motion and the local motion, we use adaptive methods respectively to deal with both kinds of the motion compensation. In our experiment, we use Cubic Curve Model to compensate the local motion of the lip part and use extended affine transform to compensate the content global motion.

Key words: moving pictures segmentation, motion compensation, Cubic Curve Model, extended affine transform

1. Introduction

To propose an efficient coding approach for moving pictures, we must consider how to segment the images and how to compensate the content motion, including the global motion and the local motion, and how to deal with the relationship between the image segmentation and the content motion compensation smoothly. We pay much attention to the head and shoulder moving pictures, such as television news or visual telephone conference. As for the static image segmentation, many literatures are published [1][2]. But for the moving pictures segmentation, the number of contents must be considered. Because a set of motion parameters is needed for each content when compensating motion. To decrease the overhead, the less number of the contents is the better. In this paper, an efficient method is proposed to control the number of the contents using motion information. Motion compensation plays an important role in moving pictures compression that exploits the high temporal redundancy between consecutive frames to obtain high compression efficiency. For a concrete content, the motion can be divide into two parts, the global motion, which is the motion of the whole content, and the local motion, which is the motion of a limited part of the content. To estimate the content motion, many methods are proposed, such as DCT-based motion estimation [3], the affine transform [4][5]. But all these methods consider only the global motion and ignore the local motion. It must decrease the compression efficiency. Some study adds the local motion as the noise into the global motion [6]. The motion parameter is a large matrix. That is to say, the overhead is very large and is not suitable for compressing moving pictures. The consideration in this paper is to separate the global motion and the local motion. It is assumed that there is only the global motion at first. We use the extended affine transform to estimate the motion. The next frame can be predicted by the motion parameters and the current frame color information. If the error of some part between the predicted image and the original image is large, it is decided that there is some local motion. In this research, we use the head-face image sequence. In the head and shoulder moving pictures,the lip motion is very important. Issue [7] gives a method for lip motion compensation. We propose a method named Cubic Curve Model [8] to compensate the lip motion.

In this research, we use the standard moving pictures named Claire. This paper is organized as follows. In Sect. 2, we explain how to segment images into contents, whose number is suitable. We describe the extended affine transform and the Cubic Curve Model in the Sect. 3. Section 4 is devoted to the discussion of the motion compensation. The experiment results are given in Sect. 5. Section 6 draws a conclusion for this paper.

2. Image Segmentation

For moving pictures coding, the number of the contents must be the least if the motions of all contents can be predicted because each content should be given a set of motion parameters. The number of contents increasing means the overhead becoming large. The number of the contents is decided by the motion information. If some part in the image moves, the difference of the pixel at the same position can be observed between two consecutive frames. Figure 1 shows one original frame in Claire. The difference from the next is changed to binary values with the threshold of 8

Fig. 1 An original frame in Sequence Claire.

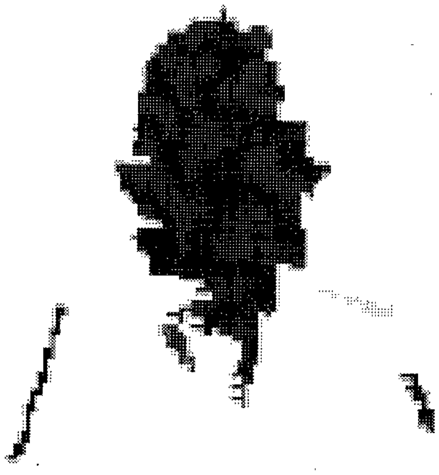


Fig. 2 Different parts between two consecutive frames after deleting noises and merging.

considering the coding method can compress the number less than 8 efficiently. Because when compressing the error data, the decimal error data should be changed into binary data with only 4 bits considering the minus data. The small point and the thin line in the difference image are regarded as the noise. We delete the noise at first. Then merge the connective pixels with large difference as one part. If there are pixels with small difference in a part, these pixels are also merged in the part. So Fig. 2 is obtained. Based on Fig. 2, the image is segmented with region growing method. A mark array is set with all the elements being 0 at first. The element of mark array at the position (0, 0) is set as 1 and get the RGB data of that pixel. If a pixel is adjacent to the pixel at (0, 0) and not in the moving part, the RGB data of this pixel is compared with those of the pixel at (0, 0). If the difference are small, the element of mark array at the same position is set as 1. Then check another unmarked pixel adjacent to the marked pixel. The process is repeated until there is no any pixel satifying the condition. So the background content is extracted. Then mark all the moving parts obtained above with number 2, 3, ... , N. That is to say, the number of contents after segmentation is N. The process extracting the moving contents is the same as that extracting the background content. After segmentation, there are maybe some pixels that can not be merged into any moving content or the background content. These pixels are merged into the adjacent content according to the distance of vector with the elements of RGB if the adjacent contents are more than one. Some pixels should be in a content but are merged into another content because the pixels are maybe in the content in the current frame but the pixels at the same position in the next frame are in another content. So we do a postprocess to deal with this problem. The pixels at the contents boundary are check again with vector distance to decide whether the pixels should be merged into the correct content. If two or more contents are adjacent and having the same attribute, here is the color, they are merged to one. For example, the head content is shown in Fig. 3.

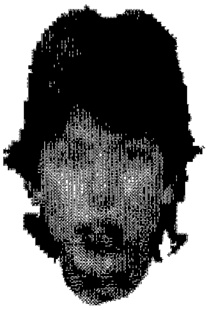


Fig. 3 The head content.

3. Extended Affine Transform and Cubic Curve Model

In the Claire sequence, all contents have the global motion, and the lip part in the head content has the local motion. We use the extended affine transform to predict the global motion. The local motion compensation of the lip part is dealt with by using the cubic curve model.

3.1 Extended Affine Transform Two dimensions affine transform is often used to predict a content global motion. But it can not describe three dimensions motion well, for example, lowering the head. We extends the two dimensions affine transform to make it be able to describe the three dimension motion but only using two dimension variables. It is shown in Fig. 4.

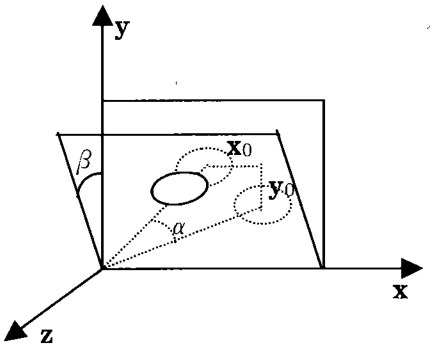


Fig. 4 Extended Affine Transform

The new position (x', y') moved from position (x, y) is given by formula (1).

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = s \begin{bmatrix} 1 & 0 \\ 0 & \cos\beta \end{bmatrix} \left(\begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \right) \quad (1)$$

The rotation in XY surface is described with α and the rotation in YZ surface is described with β . x_0 and y_0 are the translations in the direct of x, y, respectively. s is the zoom ratio.

3.2 Cubic Curve Model When predict the head content using the extended affine transform, large error occurs in the lip part because of the local motion. We predict the lip motion with the cubic curve model. The contour of the lip can be described by two cubic curves as below:

$$y = a_i x^3 + b_i x^2 + c_i x + d_i \quad (2),$$

where i = 1, 2. The lip part is contracted and approximated the contour to two cubic curves by the least square method. The position and the shape of the lip can be described only by the cubic curves coefficients. These coefficients are used in the motion compensation and the shape adaptive coding.

The cubic curves coefficients of lip parts in the two consecutive frames, and the current frame color information are used to predict the lip part in the next frame. The step is as below: 1. Allocate the pixel position named *p1* and *p2* by the two sets of motion parameters in the current and the next frame, respectively. 2. Write the color data at *p1* to *p2*. 3. Repeat Step 1 and Step 2 until all pixels in the current lip part are done.

The shape adaptive coding can be done by the cubic curves coefficients without redundant data to describe the lip shape. Because the lip contour is described by the cubic curves, the pixel number in each row can be calculated. All the pixels are relocated to a rectangle from the beginning to the end in turn that can be encoded easily. It can be turned back to the original shape losslessly when decoding.

4. Motion Compensation

From the segmentation, the original image shown in Fig. 1 is divided into five contents, the head, the left body, the right body, the center body and the background. For the background, we do no motion compensation. Because the content motions between two consecutive frames are exiguity, it is set that the rotation in XY surface is from -2 degree to 2 degree, the rotation in YZ surface is from -2 degree to 2 degree, the translations are from -2 to 2 and the zoom is from 0.5 to 1.5. All the variation steps are set as 0.1. A content in the current frame is transformed with the extended affine parameters, and the position of the content in the next frame can be got. The differences of RGB data between the current frame and the next frame are calculated. We select the extented affine parameters as the global motion parameters when the differences are the smallest.The motions of the left, right and center body contents between two consecutive frames are predicted by the extended affine transform. These three contents have almost the same motion parameters. So they are merged to one body content. With the motion parameters and the current frame color information, the body content in the next frame is predicted. The difference between the original content and the predicted content is very small. It means that the body content has no local motion, and the motion compensation can be dealt with by the global motion parameters earned by the extended affine transform sufficiently.

For the head content, the global motion parameters are shown as below. The translations should be integers, but using decimals can make the error smaller due to the rotations. The zoom ratio s is 1.0 because there is almost no zoom.

$$\begin{pmatrix} \alpha \\ \beta \\ x_0 \\ y_0 \\ s \end{pmatrix} = \begin{pmatrix} 0.8^\circ \\ 0.9^\circ \\ -1.6 \\ 0.2 \\ 1.0 \end{pmatrix}$$

The difference between the original and the predicted head content is shown in Fig. 5. It can be seen that the error is large in the lip part because of the lip part having the local motion. The average error in the lip part is about 24 and that in the other part of head is about 5. We use the cubic curve model to compensate this lip local motion. Because the lip part error is large and the color of lip part differs from that of the face, we can extracted the lip part. The extracted lip part is shown in Fig. 6 and the approximated cubic curves are shown in Fig. 7.

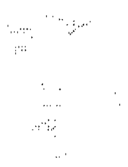


Fig. 5 Error image between the original image and the predicted image by the extended affine transform.

Fig. 6 The extracted lip part.

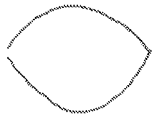


Fig. 7 Approximated cubic curvers of lip contours.

Figure 8 shows the difference between the original and predicted lip part. The error is much smaller than that before compensated by the cubic curve model.

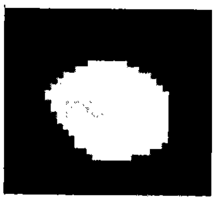


Fig. 8 Lip part error image between the original image and the predicted image by the cubic curve model.

5. Experimental Results

We choose the DCT combined with the run length Huffman coding to do the codec. The run length Huffman coding is fit for the data form with long consecutive zeros obtained by the DCT after the zigzag process. In the tables of the experimental results, N is the DCT quality factor and the DCT block size is 8. The compression ratio ***r*** is calculated with Formula 3, where ***So*** is the content original size and ***Sc*** is the size after compressing.

$$r = \frac{So}{Sc} \qquad (3)$$

The PSNR is used to express the quality of the reconstructed image, which is calculated with Formula 4.

$$RMSE = \sqrt{\frac{\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}(f(i,j)-F(i,j))^2}{M*N}}$$

$$PSNR = 20\log_{10}\left(\frac{255}{RMSE}\right) \qquad (4)$$

Where M * N is the image size, and f(i, j) and F(i, j) are the data value in position (i, j) of the reconstructed image and the original image respectively. RMSE is called root mean squared error.

The PSNR of the reconstructed body content and the compression ratio is shown in Table 1.

Table 1. The PSNR of reconstructed body content and the compression ratio.

N	PSNR	Compression Ratio
0	64.93	5.21
1	51.66	20.37
2	49.32	30.54
3	47.44	45.12
4	45.94	59.65
5	44.62	81.67

The results of compensating the head content motion with extended affine transform only are compared with the results, which compensate the local motion with the cubic curve model besides the global motion compensation. Table 2 shows the results of the lip part only. CCM is cubic curve model.

Table 2. Comparison of the results in lip part only.

	Using CCM		Without Using CCM	
N	PSNR	Compression Ratio	PSNR	Compression Ratio
0	58.75	1.65	58.75	1.47
1	43.05	5.17	41.53	4.27
2	40.10	8.32	38.73	6.46
3	38.63	11.21	36.62	8.76
4	37.65	14.71	35.39	10.92
5	37.18	18.41	34.78	12.39

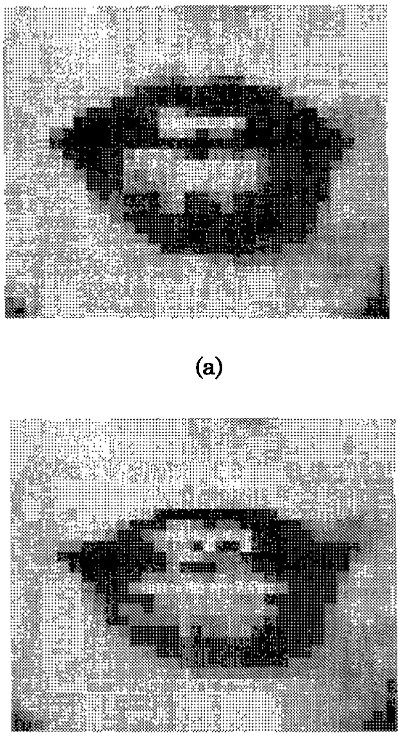
It can be seen that using CCM is better than without using

CCM in both PSNR and compression ratio. The comparison of the head content is shown in Table 3.

Table 3. Comparison of the results of the head content.

	Using CCM		Without Using CCM	
N	PSNR	Compression Ratio	PSNR	Compression Ratio
0	61.74	2.02	61.74	2.02
1	46.20	5.81	45.33	5.67
2	42.98	11.87	42.21	9.55
3	41.81	15.75	40.80	13.15
4	39.74	18.36	39.12	16.20
5	38.90	25.04	37.99	20.15

The effect of using CCM is not that obvious because the lip part is small. But from the reconstructed images, the obvious effect can be seen. In order to see clearly, we expand the lip part of the reconstructed images and show them in Fig. 9.


(a)

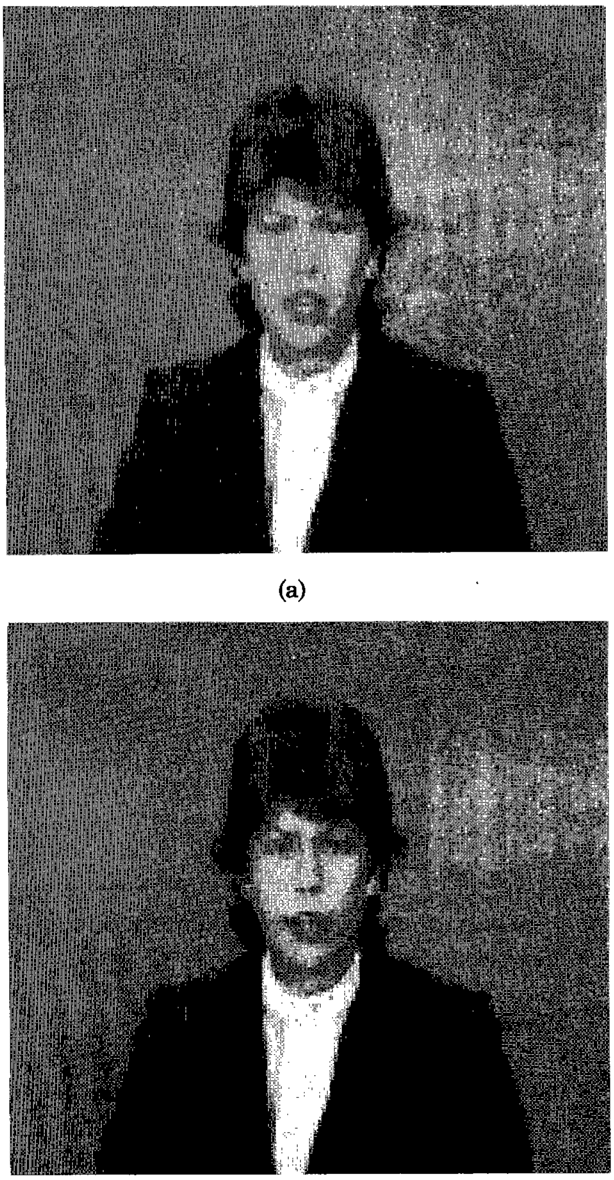
(b)

Fig. 9 Reconstructed lip part (a) Having used cubic curve model to compensate the local motion (b) Without the local motion compensation.

For the whole image, the PSNR and compression ratio is shown in Table 4, which the DCT quality factor for the head content is 2 and for the other contents is 5 and DCT block size is 16 * 16. The effect of local motion compensation with cubic curve model is not that conspicuous. But from the research of Youichi Toukura [9], it is very important for a person to grasp the information of the motion of lip part when to understand what someone is saying. The local motion compensation with cubic curve model increase both the PSNR and compression ratio obviously for the lip part. From the reconstructed images shown in Fig. 10, we can see the quality difference clearly.

Table 4. The PSNR and compression ratio of the whole image.

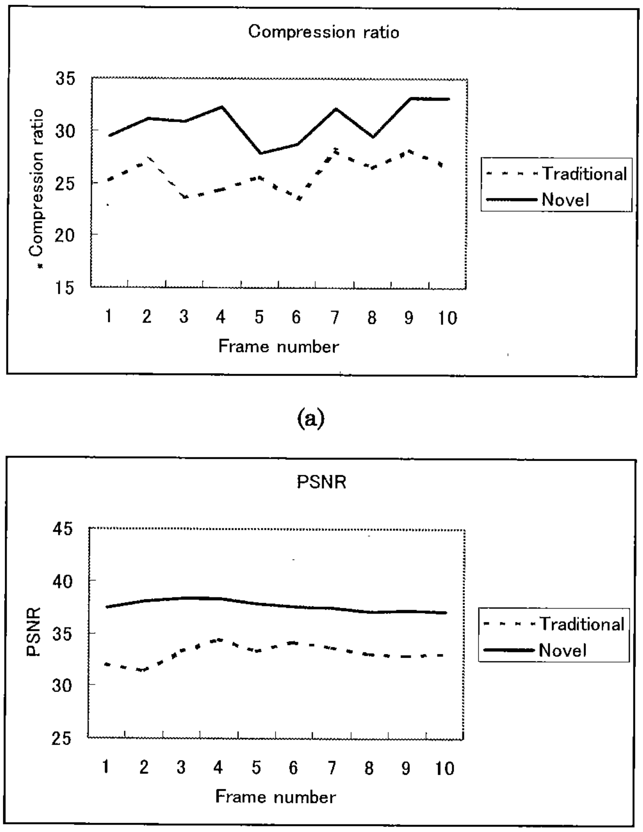
	Using CCM	Without Using CCM
PSNR	37.43	36.88
Compression Ratio	29.52	28.74


(a)

(b)

Fig. 10 Reconstructed Image (a) Compensated both the global motion and the local motion (b) Compensated only the global motion.

Traditionally, affine transform is used to compensate motion popularly [10]. We compare the compression ratio and PSNR obtained by the traditional method and those obtained by the novel method we proposed using the same moving picture Claire. We test 10 frames. The result is shown in Fig. 11. For the head content, the DCT quality factor is 2 and for the other contents it is 5. Figure 11(a) is the compression ratio result and Fig. 11(b) is the PSNR result. We can see both the compression ratio and the PSNR obtained by the method we proposed are better than those obtained by the traditional method.



(a)

(b)

Fig. 11 The result comparision of the traditional method with the novel method we proposed. (a) Compression ratio. (b) PSNR.

6. Conclusion

In this paper, we propose a simple efficient method to segment the images. It can control the number of contents to be fit for the motion compensation. It can make the overhead least. We consider the three dimension motion of the contents and extend the affine transform to make the global motion parameters much better. For the lip motion compensation, we propose the cubic curve model. From the experimental results, it is clear that this method is effective. We also compare the compression ratio and PSNR of reconstructed image obtained by the traditional affine transform and those obtained by the novel method we proposed. The novel method do much better than the traditional method.

(manuscript received March 19, 2001, re-received December 21, 2001)

References

[1] Tatsuya Yamazaki, “A Study of Color Image Segmentation Based on a Multi-dimensional Histogram”, TECHNICAL REPORT OF IEICE, pp. 9-14, OFS2000-37, IE2000-43 (2000-09).

[2] Timothy N. Jones and Dimitris N. Metaxas, “Image Segmentation Based on the Integration of Pixel Affinity and Deformable Models”, Proc. of CVPR’ 98, June 1998 Santa Barbara.

[3] Ut-Va Koc and K. J. Ray Liu, “DCT-Based Motion Estimation”, IEEE Trans. Image processing, Vol 7, No 7, pp. 948-965, July 1998.

[4] Songjun Chong and Osamu Nakamura, “Very Low Bitrate Video Coding based on a Method of Specifying Facial Area Using the Modified HSV Color System”, Trans. of the Institute of Image Information and Television Engineers, Vol 51, No. 10, pp. 1696-1705, 1997

[5] Chi-His SU, Hsueh-Ming HANG and David W. LIN, “Global Motion Parameter Extraction and Deformable Block Motion Estimation”, IEICE trans. Information and Systems, Vol. E82-D, No. 8 pp. 1210-1218, Aug. 1999.

[6] Haibo Li and Robert Forchheimer, “Two-View Facial Movement Estimation”, IEEE Trans. Circuits Syst., Vol. 4, pp. 276-287, June 1994.

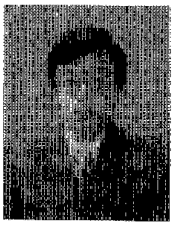
[7] F. Luthon and M. Lievin, “Lip Motion Automatic Detection”, SCIA ’97, 10th Scandinavian Conference on Image Analysis, pp. 253-260, June 1997.

[8] Zhigang Chi, Kenji Yamauchi, Toshifumi Kimura and Kennichi Hatakeyama, “Motion Compensation and Shape Adaptive Coding of Moving Pictures by Cubic Curve Model”, T. IEE Japan, Vol. 120-C, No. 10, pp. 1502-1503, 2000

[9] 東倉　洋一，“人とコンピュータを結ぶ科学と技術”，東京情報大学研究論集，Vol. 2, No. 3, (1998.4) pp. 179-185.

[10] Ken Ito and Shigeyuki Sakane, "Visual Tracking Based on Dynamic Transition in Groups of Affin Transformed Templates", Journal of the Robotics Society of Japan, Vol. 19, No. 1, pp. 100-108, 2001.

Zhigang Chi



(Non-member) was born in Harbin, China. He received B.S. degree from Beijing University of Posts and Telecommunications, China in 1995 and M.S. degree from Himeji Institute of Technology, Japan in 1999. He is a doctor condidate in Himeji Institute of Technology. His major research is on image processing.

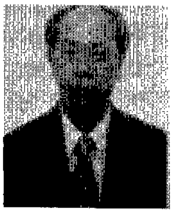
Toshifumi Kimura

(Non-member) was born in Hyogo, Japan. He received the B.S. and M.S. degrees from Himeji Institute ofTechnology, Hyogo, Japan, in 1994 and 1996, respectively. In 1998, he joined the School of Humanities for Environmental Policy and Technology, Himeji Institute of Technology as a research associate. His main research area is image processing.

Kenji Yamauchi

(Member) was born in Hyogo, Japan, in 1941. He graduated from Himeji Institute of Technology, Japan, in 1963 and received the Master and Doctor of Engineering degrees from Osaka University in 1967 and 1970. Since 1970, he has been with the Himeji Institute of Technology and ia currently a Professor and doctoral superviser in Department of Electronics Engineering. He joined the technical committee of electromagnetic compatibility in Japan as a member from 1985 to 1998. His main interest is the study of intelligent image processing.

Kennichi Hatakeyama



(Non-member) received the B.E., M.E, and Ph.D. degrees in electrical engineering from Tokyo Metropolitan University, Tokyo, Japan, in 1976, 1978, and 1992, respectively. From 1978 to 1997, he has worked at NEC Corporation, Japan, where he was engaged in the research and development of absorbing/shielding material and anechoic chamber design. He is currently an associated professor of Himeji Institute of Technology, Japan.