Journal of Multimedia and Information System Vol. 2, No 1, March 2015(pp. 163-170): ISSN 2383-7632(Online)
http://dx.doi.org/10.9717/jmis.2015.2.1.163

Automatic Name Line Detection for Person Indexing Based on Overlay Text

Sanghee Lee[1], Jungil Ahn[2], and Kanghyun Jo[3,*]

Abstract

Many overlay texts are artificially superimposed on the broadcasting videos by humans. These texts provide additional information to the audiovisual content. Especially, the overlay text in news videos contains concise and direct description of the content. Therefore, it is most reliable clue for constructing a news video indexing system. To make the automatic person indexing of interview video in the TV news program, this paper proposes the method to only detect the name text line among the whole overlay texts in one frame. The experimental results on Korean television news videos show that the proposed framework efficiently detects the overlaid name text line.

Key Words: automatic indexing, name detection, news video, overlay text, person identification.

I. INTRODUCTION

Nowadays, many overlaid graphics such as text and channel logos are artificially superimposed on the broadcasting videos by human producers. The graphic text inserted videos is called the overlay text. This text differs from the scene text which naturally occurs in the scene being recorded such as advertising boards, street signs and clothing. The overlay text provides additional information from the audiovisual context so the audience pays more attention. For this, it is superimposed on the video frame during the editing stage of production. Therefore, the extraction of video text information has a very important significance for the further semantic understanding. Many approaches have been studied in the scene understanding, indexing, browsing, and retrieval [1-5].

Especially, the overlay text in news videos provides concise and direct description of the content. For instance, the text annotates the names of people and places, or describes objects and the current issue.

Therefore, the overlay text is the most reliable clue for constructing a news video indexing system, when the text can be accurately transcribed. The detection and recognition of the overlay text have become a hot topic in news video analysis, such as identification of person or place, name of newsworthy event, date of event, stock market, other news statistics, and news summaries [6-9].

Among these applications, the identification of the person from the overlaid text raises a lot of interest in the information research community. The identification using the overlaid person names (OPN) has started to be investigated [10]. Since then the research area has raised a large amount of work, especially in face clustering tasks, face naming of captioned images, and recently, automatic naming within broadcast videos [11-16].

However, this paper focuses on the application to make the automatic person indexing system by the OPN in the news interview videos. The name and title information in the interview video of the TV news program are valuable for building an information retrieval and data mining system.

As the first step for this goal, this paper proposes the method to only detect the name text line among the

Manuscript received March 26, 2015; Revised April 10, 2015; Accepted April 20, 2015. (ID No. JMIS-2015-0008)
Corresponding Author(*): Kanghyun Jo, The University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan City, Korea 680-749, Tel: +82-52-259-2208, E-mail: acejo@ulsan.ac.kr
[1,3]Electrical Engineering, The University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan City, Korea, E-mail: tusun49@gmail.com, acejo@ulsan.ac.kr,
[2]Ulsan Broadcasting Corporation, 41 Kugyo-ro, Jung-gu, Ulsan city, Korea, E-mail:

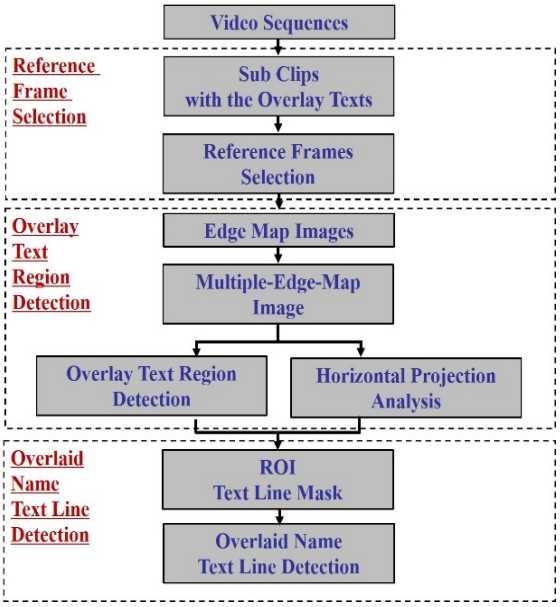


Fig. 1. The framework of proposed method.

whole overlay texts in the one frame. The framework of the proposed method is given in Figure 1. The paper's method uses the rule-based characteristics in production of the TV news program. Many of the accepted production rules apply to the TV content, since the broadcasting videos are produced by professionals. For example, text and logo are often overlaid onto the natural content in a structured manner, such as aligned text lines at the bottom or on the upper corners of the screen, to minimize the chance of covering the important content.

This system contains three main parts: the reference frame selection, the overlay text region detection, and the overlaid name text line detection. The proposed method is explained in detail in the following section. Section III shows the results of the overlaid name text line detection and describes the analysis of the experimental results. And the last section describes the brief conclusion and the future works.

II. PROPOSED FRAMEWORK

For the videos have the appearance and disappearance of the overlay text, the processing of all frames is time-wasting. As the first step, the sub clips which contain the overlay texts are made based on the corner characteristics of the text. Second, to reduce the processing time, the reference frames are selected. Because the same overlay text lasts on the same position for a few seconds or more. The third step transforms the reference frames into grayscale images. Next step does the logical AND operation on the edge map images. And based on this

(a) Frame 1 (b) Frame 15 (c) Frame 65

Fig. 2. The example of overlay text appearance in interview video.

result, the overlay text region is obtained from the number of black and white transitions and also the horizontal projection histogram is acquired. And then to limit the region of interest in the whole text lines, the detected overlay text region and the horizontal projection analysis is combined. At last, the overlaid name text line is detected by overlaying the ROI text line mask with one of the four reference frames. Details are given below.

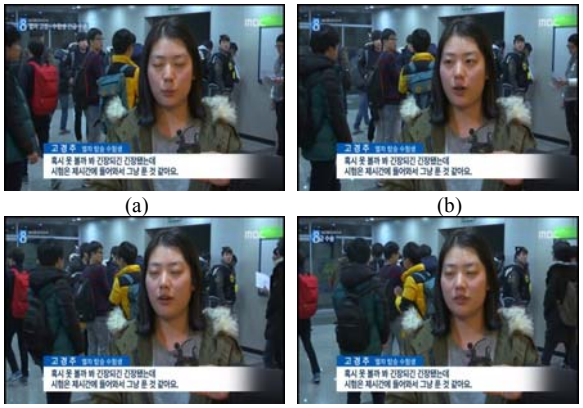
1. Reference Frame Selection

By observing a large quantity of the TV news programs, the overlay text superimposed on most news videos has the following characteristics. Since the appearances and disappearances of the overlay text occur suddenly or slowly like Fig. 2, all frames in the video sequences do not have the overlay text. The position of the overlay text is fixed; generally in the range of 1/2 from the bottom of the frame. The overlay text is aligned horizontally and overlaid on the opaque or translucent background matte. For readability in a complex scene, the same overlay text appears in the same position for a few seconds or more.

First, when the video sequences come in, the frame images are stored in RGB color space. And then, according to the characteristics presented above, this paper decides whether the overlay text is included or not in all frames of the videos, and yields the sub clips with the overlay texts. For, the detection of the overlay text in every frame is time-wasting. Therefore, to decide the frames included the overlay text, this paper uses the corner density map based on Harris corner detector proposed in the previous work [17].

And next, the reference frames are selected in the sub clips by the method proposed in [7]. If videos are played f frames per second, the overlay text stays in a fixed location for at least $2f$ consecutive frames. Let k be the nearest integer that is not less than f. This paper defines every consecutive k frames to be one round. To simplify the calculation, about only 1st round, the four reference frames are selected on frame 1, $\lfloor k/3 \rfloor$, $2\lfloor k/3 \rfloor$, $3\lfloor k/3 \rfloor$ like Fig. 3. Because the same overlay text is fixed in the same position for every consecutive k frames.

2. Overlay Text Region Detection



(a) (b) (c) (d)

Fig. 3. The four color reference frame images. (a) 1st reference frame. (b) 2nd reference frame. (c) 3rd reference frame (d) 4th reference frame.

(a) (b) (c) (d)

Fig. 4. The four gray-scale reference frame images. (a) 1st reference frame. (b) 2nd reference frame. (c) 3rd reference frame. (d) 4th reference frame.

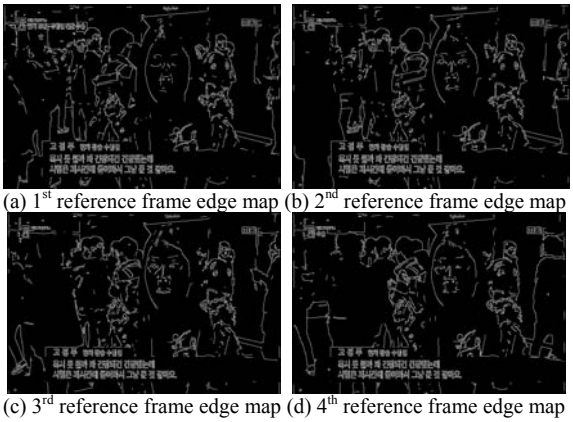
Next step is to detect the whole overlay texts in the frame by using the four reference frames of the previous result. This paper uses the temporal information of video and the logical AND operation on Canny edge maps to detect the overlay text region [7].

First, this system transforms the four color reference frames into grayscale images by (1)

$$\mathrm{Y} = 0.299\mathrm{R} + 0.587\mathrm{G} + 0.114\mathrm{G} \qquad (1)$$

where Y is the intensity value and R, G, B are the value of red, green, blue channel of the pixel, respectively. Fig 4. shows the conversion result.

Second, this system yields the edge map images by the Canny edge detector and the simple line deletion applied on each of the grayscale images. The simple line deletion is used to remove long lines which are unlikely to be characters in the Canny edge result image. When the Canny edge image is scanned from left to right and top to bottom, a horizontal line vertical line is removed if its length exceeds the presumed width w and height h of a character. As a result, the edge map images are obtained as shown in the Figure 5.



(a) 1st reference frame edge map (b) 2nd reference frame edge map (c) 3rd reference frame edge map (d) 4th reference frame edge map

Fig. 5. The four gray-scale reference frame edge images.

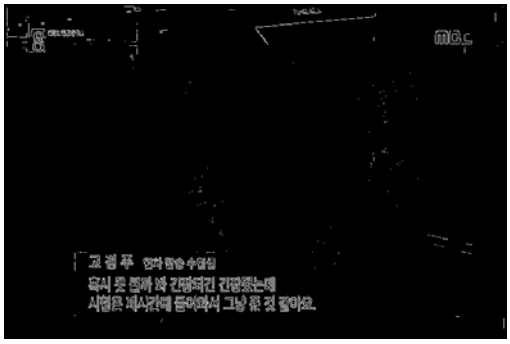
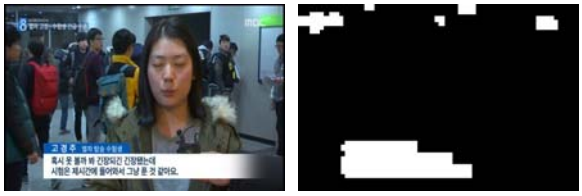


Fig. 6. The Multiple-Edge-Map image.

Third, the logical AND operation on four Canny edge map images is executed. The result image is called the Multiple-Edge-Map. After the AND operation, a position (i, j) becomes an edge pixel if all four edge images are edge at (i, j). Therefore, most of the background edge pixels are removed, whereas the static overlay texts are remained. Because, the same overlay text appears in the same location for many successive frames, while the location of background edge pixels may differ in a few pixels. The Figure 6 shows the Multiple-Edge-Map image. The result well explains that the problem which the difficulty to distinguish whether the detected edges are really from overlay texts is alleviated by multiple frame integration method.

Fourth, the overlay text candidate region is detected by utilizing the number of the black and white transition. As shown in (2), the value of N_{trans} can be obtained that a window of the presumed character size $w \times h$ slides from left to right and top to bottom on the Multiple-Edge-Map image.



(a) (b)

Fig. 7. The overlay text region detection image. (a) 1st reference frame. (b) Overlay text region.

(a) (b)

Fig. 8. The horizontal projection image. (a) Multiple-Edge-Map image. (b) Horizontal projection image

$$N_{trans} = \sum_{i=0}^{h-1}\left(\sum_{j=1}^{w-1}\left|b(i,j)-b(i,j-1)\right|\right) + \sum_{j=0}^{w-1}\left(\sum_{i=1}^{h-1}\left|b(i,j)-b(i-1,j)\right|\right) \quad (2)$$

where w and h are the width and height of window, and $b(\cdot)$ is binary image. If N_{trans} is larger than threshold T_{trans}, this window is masked. The union of all masked windows is the overlay text candidate region. The threshold T_{trans} depends on the character size and is obtained by $T_{trans} = \beta(w \times h)$ with β a constant which is empirically measured.

Finally, to resolve the problem that characters lose some pixels in the AND operation, a morphological closing is applied first and then dilation is followed. A closing with a horizontal structuring element of size $\lceil w/3 \rceil$ is used to fill holes. A dilation with a structuring element of size $\lceil w/4 \rceil \times \lceil h/4 \rceil$ is applied to the connect characters. The result image is the overlay text region like Figure 7.

3. Overlaid Name Text Line Detection

In general, many overlay texts can exist in the one frame of the video. To detect a name text line, it is necessary not analyzing the whole overlay texts in the one frame. This paper constrains the detection region based on the news program production rules. The TV content is produced by professionals. Many of the accepted production rules apply to the TV content. In the news interview video sequences, over a few lines, the story of interviewees is positioned at the bottom of the frame like the characteristics remarked in the previous phrase. And the interviewee's name and the title fix on the top line among the interviewees first story lines and appear in the same position for a few seconds. Therefore, only the top of the first story lines is the interest of region (ROI).

(a) ROI text line mask (b) Overlaid name text line image

Fig. 9. The overlaid name text line. (a) ROI text line mask. (b) Overlaid name text line image.

To detect the ROI text line, at first, the horizontal projection histogram must be obtained by applying to the Multiple-Edge-Map image of the previous phrase. To scan the result image from top to bottom, and count the number of edge in a row can be gotten the horizontal projection histogram image like (3).

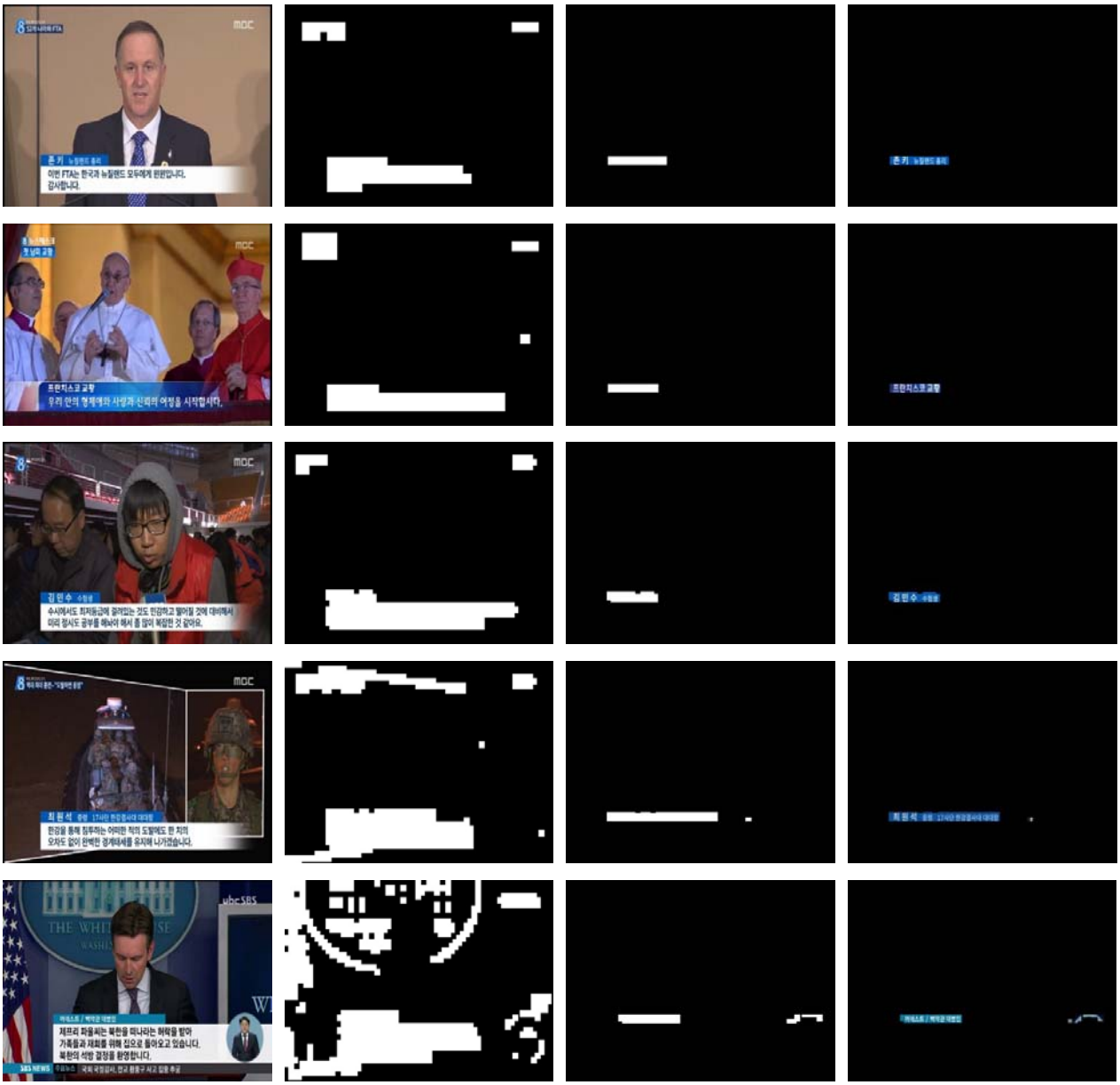
$$H_{hor}(i) = \sum_{j=0}^{w-1} b(i,j), \quad i = 0,1,\cdots,h-1 \quad (3)$$

The Figure 8 shows the result image. This projection image is analyzed in the range of 1/2 from the bottom of the frame. The first top area of the horizontal projection histogram in the half bottom region is selected as the region of interest. The start point and end point of ROI along the height (vertical) axis are applied to the overlay text region image. The result is the ROI text line mask image like Figure 9(a). At last, to apply to the one frame of four reference frame images yields the overlaid name text line image such as Figure 9(b).

III. EXPERIMENTAL RESULTS AND ANALYSIS

Since there is no standard dataset for overlay text in videos, the test videos used in the experiment were captured from the interview video sequences in a TV news program in Korea. The resolution of the videos was 720×480, and frame rate was 29.97 frames per second. The presumed character size $w \times h$ was 20×20 pixels. The threshold of the N_{trans} was set to be 0.15 and the threshold of the horizontal projection histogram analysis was the presumed character width 20.

Figure 10 and table 1 show the results of the overlaid name text line detection. To evaluate the performance of the proposed algorithm, this paper shows the block level accuracy of the results of overlaid name text line detection, and uses precision and recall as performance measures. TP (true positive) is the predicted positive



(a) (b) (c) (d)

Fig. 10. The example images of the experimental result. (a) original frame image. (b) overlay text region. (c) ROI name line mask. (d) overlaid name text line.

block and FN (false negative) is the predicted negative block of the actual name text line. FP (false positive) is the predicted positive block and TN (true negative) is the predicted negative block of the actual non-name text line. The measures are calculated (4) and (5), and the result is shown in Table 1.

Table 1. Experimental results of name text line detection.

Method	Precision rate (%)	Recall rate (%)
Proposed method	66.67	100

$$R_{pre} = \frac{TP}{TP+FP} \quad (4)$$

$$R_{rec} = \frac{TP}{TP+FN} \quad (5)$$

In Figure 10, (b) column is the results of the overlaid text region, (c) column is the results of the ROI name line mask based on horizontal projection analysis, (d) column is the results of the detected overlay name text line. The proposed method accurately detects the position of the overlaid name text line in many examples. However, the fourth and the fifth row results show that the false block in overlay text name line was

detected. In this case, many edges have in non-text region and this area is detected as text region. Thus, it is necessary to refine process to remove this noise.

In general, the name text line contains not only name but also age, degree, job and address. This kind of information plays an important role to develop an automatic person indexing system. Therefore, it more necessary exactly detects of the text line for the input of the traditional OCR (Optical Character Recognition).

IV. CONCLUSION

This paper proposes the detection method of the name text line among many overlay texts in one frame for automatic person indexing of the interview videos. The overlay text region is obtained by edge and multiple frame integration method. And the overlaid name text line is detected by the production rules and the horizontal projection histogram analysis.

The result image is used as the input, which recognizes the text with the OCR. As a result, the retrieval system to control effectively and automatically mass the interview videos in the news can be developed.

REFERENCES

[1] Z. Wang, L. uang, X. Wu, and Y. Zhang, "A survey on video caption extraction technology", in *Fourth International Conference on Multimedia Information Networking and Security*, Nov. 2012.

[2] P. Shivakumara, TQ. Phan, CL. Tan Hong, and K. J. Lim, "A gradient difference based technique for video text detection", *in ICDAR*, pp. 15-160, 2009.

[3] U. Gargi, S. Antani, and RE. Woods, "Indexing text events in digital video database", *Pattern Recognition*, vol. 1, pp.1481-1483, 1998.

[4] P. Shivakumara, W. Huang, and CL. Tan, "An efficient edge based technique for text detection in video frames", in *DAS*, pp. 307-314, 2008.

[5] F. Xiaoling and G. Hua, "Gray-based news video text extraction approach", in *5th International Conference on computer Science and Convergence Information Technology*, 2010.

[6] Z. Yang and P. Shi, "Caption detection and text recognition in news video", *in 5th International Congress on Image and Signal Processing*, 2012.

[7] S. Huey, H. W. Chang, C. J. Wang, and C. W. Wang, "Robust news video text detection based on edges and line-deletion", *WSEAS Transaction on Signal Processing*, vol. 6, no.4, October 2010.

[8] T. Sato, T. Kanade, E. K. Huges, M. A. Smith, and S. Sato, "Video OCR: Indexing digital news libraries by recognition of superimposed caption", *Multimedia Systems*, vol. 7, no. 5, pp.385-395, January 1999.

[9] J. Poignant, L. Besacier, G. Quenot, and F. Thollard, "From text detection in videos to person identification", in *International Conference on Multimedia and Expo*, 2012.

[10] S. Satoh, Y. Nakamura, T. Kanade, "Name-It: Naming and detecting faces in news videos", *Proc. of IEEE Multimedia*, 1999.

[11] P. Gay, G. Dupuy, C. Lailler, J. Odobez, S. Meignier, and P. Deleglise, "Comparison of two methods for unsupervised person identification in TV shows", *in 12th international workshop on content based multimedia indexing*, 2014.

[12] P. T. Pham, T. Tuytelaars, and M. Mones, "Naming people in news videos with label propagation", *in Proc. of ICME*, 2010.

[13] B. Jou, H. Li, G. Ellis, D. Morozoff-Abegauz, and S. F. Chang, "Structured exploration of who, what, when, and where in heterogeneous multimedia news source", *in Proc. of ACM Multimedia*, 2013.

[14] J. Poignant, L. Besacier, V. B. Le, S. Rosset, and G. Quenot, "Unsupervised speaker identification in TV broadcast based on written names", *in Proc. of Interspeech*, 2013.

[15] J. Poignant, H. Bredin, V. B. Le, L. Besacier, C. Barras, and G. Quenot, "Unsupervised speaker identification using overlay texts in TV broadcast", *in Proc. of Interspeech*, 2012.

[16] M. Bendris, B.Favre, D. Charlet, G. Damnati, G. Senay, R. Auguste, and J. Martinet, "Unsupervised face identification in TV content using audio-visual sources", *in Proc. of CBMI* 2013.

[17] S. Lee, H. Park, J. Ahn, Y. On, and K. Jo, "Overlay text graphic region extraction for video quality enhancement", *JBE*, vol. 18, no. 4, pp-559-571, July 2013.

Journal of Multimedia and Information System Vol. 2, No 1, March 2015(pp. 163-170): ISSN 2383-7632(Online)
http://dx.doi.org/10.9717/jmis.2015.2.1.163

Authors

Sanghee Lee received the B.E. degree in electronic engineering from the Kyoung-book National University, Daegu, Korea, in 1994, and M.E. degree in Institute of e-Vehicle Technology, University of Ulsan, Korea, in 2009.

He is currently pursuing the Ph.D. degree in school of electrical engineering from the University of Ulsan, Korea. He is working at the Ulsan Broadcasting Corporation, Ulsan, Korea as a broadcasting engineer since 1997. His research interests are in the area of multimedia broadcasting, indexing and retrieval sys tem, intelligent system, and computer vision.

Jungil Ahn received the B.E. and M.E. degrees in electronic engineering in the Kyoungbook National University, Daegu, Korea, in 1986, and 1988, respectively.

He is currently pursuing the Ph.D degree of in school of electrical engineering from the Kyoungbook National University, Daegu, Korea. He is working at the Ulsan Broadcasting Corporation, Ulsan, Korea as a broadcasting engineer since 1997. His research interests are in the area of multimedia broadcasting, indexing and retrieval sys tem, intelligent system, and computer vision.

Kanghyun Jo received his B.E. degree in Mechanical and Precision Eng. From Busan National University, Korea and his M.E. and Ph.D. degrees in Computer-Controlled Machinery Eng. from Osaka University, Japan, in 1989, 1993, and1997, respectively.

He is currently a Professor at the Faculty of Electrical Eng. and Information Systems, University of Ulsan, Korea. His research interests include computer vision, human-computer interaction, robot applications in town and health-care and intensive intelligent systems. He is actively participating as a member of in very many professional research societies, like IEEE, IEEK, ICROS, KRS, KIPS, KIISE, and KSAE.

This is blank Page