International Journal of Fuzzy Logic and Intelligent Systems, vol. 12, no. 4, December 2012, pp. 308−318
http:// dx.doi.org/ 10.5391/IJFIS.2012.12.4.308 PISSN 1598−2645 eISSN 2093−744X

Size-Independent Caption Extraction for Korean Captions with Edge Connected Components

Je-Hee Jung[1], Jaekwang Kim[1]and Jee-Hyong Lee[1]

[1]Department of Electrical and Computer Engineering, Sungkyunkwan University
2066 Seobu-ro, Jangan-gu, Suwon-si, Gyeonggi-do 440-746, Republic of Korea

Abstract

Captions include information which relates to the images. In order to obtain the information in the captions, text extraction methods from images have been developed. However, most existing methods can be applied to captions with a fixed height or stroke width using fixed pixel-size or block-size operators which are derived from morphological supposition. We propose an edge connected components based method that can extract Korean captions that are composed of various sizes and fonts. We analyze the properties of edge connected components embedding captions and build a decision tree which discriminates edge connected components which include captions from ones which do not. The images for the experiment are collected from broadcast programs such as documentaries and news programs which include captions with various heights and fonts. We evaluate our proposed method by comparing the performance of the latent caption area extraction. The experiment shows that the proposed method can efficiently extract various sizes of Korean captions.

Keywords: Caption extraction, Edge connected components, Korean caption, Size-independent.

1. Introduction

Due to the development of information technology, people easily create, store, search and transmit the data which contains various contents. The content is divided into simple contents, such as text and music, and multimedia contents which consist of image, sound and text such as video. Especially, the volume of video is significantly increasing, and thus the need for management of video contents is also increasing; such as searching, classification, summarization, etc.

For the efficient management of videos, information on video is crucial. The methods to obtain the information are divided into using the tag information and extracting features from video data. The former is simple and fast; however, it can be easily influenced by human knowledge. The latter is called a content based search method. The methods search video by complex computation; however, it can also search video without tag information.

The contents based video search methods include one method which extracts the captions for searching the video. The captions that appear in images are important indicators while searching for information of an image. Captions are composed of some texts and created in order to efficiently convey additional videos information. Therefore, caption extraction research has been developed to extract headlines of news program, and the names of players or sports teams.

The captions exist with various font types and sizes. To extract captions which include various font types and sizes, the existing caption extraction research analyzes the caption then proposes their own methods on their assumption of features of caption [1][2]. However, most of the existing caption extraction methods only focus on English captions and do not focus on some other languages. The English captions consist of English characters and Arabic numbers. Arabic numbers and English characters consist of one stroke or two strokes. However, Korean characters consist of two strokes at least. The stroke density of Korean is higher than English. Since texts may have different characteristics depending on the language, the existing English caption extraction methods may not guarantee that the method is correctly applied to Korean captions. Especially, caption size is a very important feature because the caption size influences morphological features of caption such as ‘stroke width’ and ‘stroke density’ in the same size location. So, we propose a size-variable Korean caption extraction method using Korean language dependent features as well as language independent features. Then, we evaluate the caption extraction method by each caption size.

The proposed method consists of five steps. The first step is called a preprocessing step which converts a colored input image into a gray image and removes noise pixels to reduce processing time and error. In order to find the existing location of a caption in an image, the second step which is called the caption location extraction step finds the representative colors of captions by color clustering. The third step extracts the

Manuscript received Nov. 29, 2012; revised Dec. 19, 2012; accepted Dec. 24, 2012.
*Corresponding author: Jee-Hyong Lee (jhlee@ece.skku.ac.kr)
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.2012-008062)
This work was supported by the IT R&D program of MKE/KEIT [KI001810041244, SmartTV 2.0 Software Platform]

edges which are included in caption existing location by checking each edge pixel. The fourth step groups the edges into connected components (CCs) and classifies CCs into caption CCs and non-caption CCs then, the location of the caption CC is extracted for a local threshold in the final step. The final step makes a binary image which includes caption pixels as the commercial OCR engines.

We evaluate our proposed method by comparing the performance of the latent caption area extraction. The experiment shows that the proposed method can efficiently extract various sizes of Korean captions.

The rest of this paper is organized as follows: Section 2 reviews the related work. Section 3 describes language dependent caption features and Korean caption features. Then, Section 4 describes our proposed method and Section 5 presents and discusses the experimental results. Finally, the conclusion is presented in Section 6.

2. Related Work

In order to extract captions from images, we need to know the features of captions. Therefore, the existing methods make a survey on caption features for caption extraction. Lyu classifies features of caption into two categories; language dependent features and independent features [1]. The language dependent features are stroke density, font size, aspect ratio, and stroke statistics. The language independent features are contrast, color, orientation, and stationary locations. The features of caption include caption's height, arrangement, inter-character distance, color, motion, boundaries, etc. [3]. Most of research on caption extraction is based on language independent features for English. Since texts may have different characteristic dependently on the language, the existing English caption extraction methods may not guarantee that the method is correctly applied to Korean caption.

Most related work was composed of three steps: text detection, text localization and text segmentation. The three step methods reduce pixels which were not included caption. Text detection step detects the part of images which include captions to accelerate the second step. Text localization step searches for the location of each character of caption. Text segmentation step separates the caption pixels from its location.

One of the existing methods for caption extraction is block-wise edge detection [4][5]. Chun's et al. uses neural networks for detecting caption existing location [4]. Wong's et al. uses line detection for the same purposes [5]. These two methods are limited. Since, their methods use fixed masks which are derived by their experiments. Their performances are dependent on the block size. Others proposed methods are morphology-based [2][6-10]. Initially, edge pixels are detected by a morphological element. Next, groups of edge pixels are merged with neighboring groups. Then, the groups which are less possibly included in caption are removed. This method's caption extraction performance is dependent on the elements of the morphological operator. Therefore, this method cannot guarantee the extraction of various size captions.

Ryu et al. [1] propose a sequential multi-resolution method for extracting variously sized captions. The method performs the same process as the mentioned ones, but multiply applies the process to the source image converted in various resolutions. Thus, the processing time increases and it cannot be guaranteed that Korean captions is property extracted because the methods extract the caption location by their experiments of English and Chinese. Korean syllables contain much strokes than English, thus the shape in lower resolution and higher resolution may be quite different.

Jeong et al. propose a stroke-based edge detection method [2]. The algorithm assumes that the thickness of a stroke is fixed and all the strokes have the same thickness. So, the method is not applicable if the thickness of strokes is not the same or larger than the fixed value.

Most existing methods focus on English characters, and they are based on morphological operations. However, it is not guaranteed that morphological operations work correctly with other languages and various size captions. Especially, Korean is composed with more various strokes and shapes. Therefore, we propose a method that robustly extracts various sized Korean captions.

3. Caption Features

In this section, we will describe language independent features and Korean specific features of caption.

3.1. Language Independent Caption Features

Most captions have some features in common regardless of language because caption is usually presented in a way that people can easily read it. The characters in captions usually lie horizontally in a uniform color. Sometimes they can appear as non-planar texts with special effects. So we assume that captions lie horizontally in a uniform color and the color of captions is brighter than background color.

Since the caption characters usually have a different color from background, the edges of characters are connected to each other and form a closed loop, i.e. a connected component (CC). Thus, all the edges of a caption character belong to the same CC. There is the corresponding edge for an edge in a CC. The corresponding edge of an edge is the mirror image reflected to the line perpendicular to the edge direction. For example, Figure1 shows outside edges of 'L'. The arrow in edge pixel is the direction of the edge. Those edges of 'L' belong to the same CC and each edge has its corresponding edge. For example, edge A and edge B are mirrored and thus corresponding to each other. In Figure1, boxes are pixels in image. Black boxes are background pixels and white boxes are caption pixels. Arrow means direction of edges.

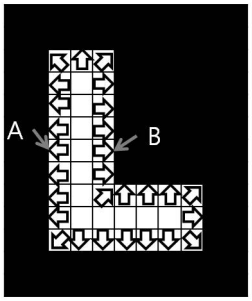


Figure 1. All Edges of character 'L'

Although captions are designed in a single color, the colors of caption pixels in an image can be different a little because of image compression. Since, images are usually compressed into image standard formats such as jpg. Especially, if characters in captions are as small as 10 pixels then the color of caption may easily be blurred by compression.

However, some pixels in image are less influenced than others by image compression. These are stroke center pixels as shown in Figure 2. Figure 2 and Figure 3 illustrate horizontal and vertical center pixels which exist on between 3:00 direction and 9:00 direction edges and 6:00 direction and 12:00 direction edges respectively. Gray boxes are the horizontal and vertical center pixels.

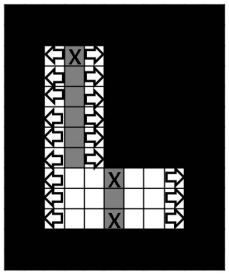


Figure 2. Horizontal center pixels

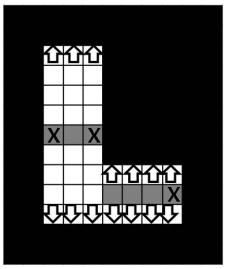


Figure 3. Vertical center pixels

Since the stroke center pixels are less influenced by image compression less than other pixels. However, some stroke center pixels which are marked by 'x' shown in Figure 2 and Figure 3. Therefore, we define cross center pixels as a pixel which includes horizontal center pixel and vertical center pixel. The cross center pixels position inner of stroke, the color of cross pixels are close to the color of captions. Figure 4 illustrates the cross center pixels as grayed boxes.

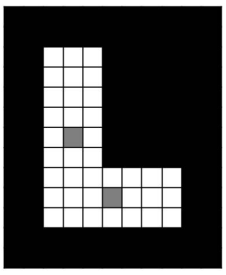


Figure 4. Cross center pixels

3.2. Korean Dependent Caption Features

We describe Korean caption features in this section. Korean characters are written by syllables which a combination of consonants and a vowel.

The vowels are shown in Figure 5, 6, and 7. Korean vowels can be classified into three types: vertical, horizontal and double vowels [11]. The vertical vowels are tall and narrow as Figure 5. The horizontal vowels are wide and short as Figure 6. The double vowels are a combination of horizontal vowels and vertical vowels as Figure 7. The consonant of Korean can classify single consonant or double consonants. Single consonant is one consonant of Korean as Figure 8 and double consonant consists of two consonants as Figure 9. For example, "갈" is a syllable in Korean, the combination of a consonant of "ㄱ", or "ㄱ" ("ㄱ" and "ㄱ" are regarded as the same) on the top left, a vowel of "ㅏ" on the top right and "ㄹ" on the bottom. Another examples are "대", "한", "민", "국", etc. Usually, a Korean syllable has one vowel and one or two consonants.

Figure 5. Vertical vowels

Figure 6. Horizontal vowels

Figure 7. Double vowels

Figure 8. Single consonants of Korean character

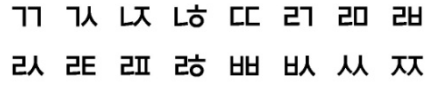


Figure 9. Double consonants of Korean character

The Korean syllables are combined with three major rules and three minor rules. If you have an interest in the rules, you may refer to [11].

The feature of Korean caption is influenced by Korean syllables. Since Korean syllables always include one vowel and one or two consonants, we can know that Korean text includes many vertical and horizontal edges. Therefore, one of Korean caption feature is that most edges of Korean are 12:00, 9:00, 3:00 or 6:00 direction. Especially, 6:00 and 12:00 direction exist on horizontal lines with other by neighbor shown in Figure 10. Figure 10 shows "우리" that is some character of Korean. White boxes are the pixels which are included caption,

but black boxes are the pixels which are not. And arrow is edge direction of 6:00 or 12:00.

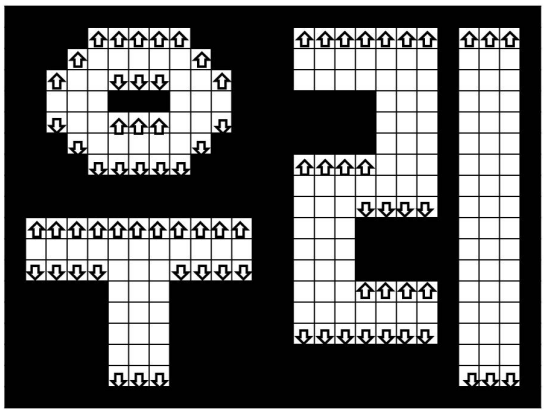


Figure 10. 6:00, 12:00 direction edges of "우리"

4. Size-Independent Caption Extraction

In this section, we will describe the proposed algorithm which is a size-independent caption extraction.

4.1. Overview

The proposed method consists of five steps as shown in Figure 11. The input image is an RGB colored image. First, we perform the color and noise reduction for fast caption extraction processing in Preprocessing step. In order to extract caption location which includes caption pixels, we check whether color of pixels is similar to the color of captions or not in Location Extraction step. All pixels in a character of captions are uniform color ideally. However, the colors of pixels are not uniform by image compression. So, we extract candidate caption location which may include captions by clustering of representative colors of captions. However, caption candidate location includes not only edges of caption but also edges of non-caption because some background pixels may have colors similar to caption. Therefore, we check the features of extracted edges in Caption Edge Extraction step for classification of pixels. In order to remove edges of non-caption, we make edge connected components (ECC) and extract 10 features of CCs to identify ECCs by decision tree in Caption CC Extraction step. Since, captions with wide stroke such as width of stroke includes 5 pixel-distances do not contain edges in stroke center. Therefore, we segment pixels of captions by threshold of gray value of pixels in Text Segmentation step. Most boundary edges exist between pixels of caption and non-caption. So, we use boundary gray values of edge pixel to determine the gray value of threshold whether pixels is caption pixels or non-caption pixels. The text image which will be used as the Optical Character Reader (OCR) input for character recognition is the final output.

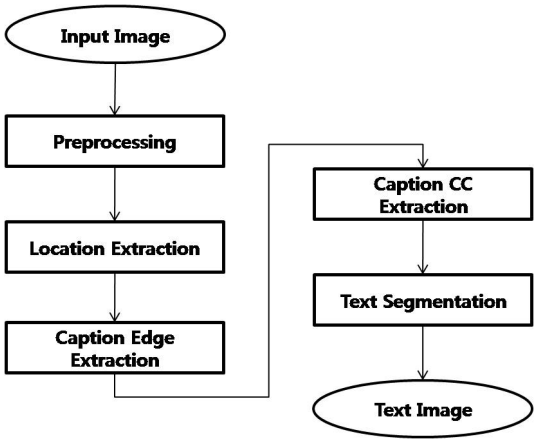


Figure 11. Overview of caption extraction method

4.2. Preprocessing

The goal of preprocessing step is to convert the input image into a gray image to decrease processing time and to remove noisy pixel for error reduction. We convert the input image into a gray image by color reduction. Since the input image includes three channels: red, green and blue colors, a colored input image '*C*' can be converted into a gray image '*G*' by (1).

$$G(x,y) = 0.3 \times C_R(x,y) + 0.59 \times C_G(x,y) + 0.11 \times C_B(x,y) \quad (1)$$

where x and y are pixel positions in the image. C_R, C_G and C_B are red, green, and blue channels of the image. G is the gray scale image that contains 256 gray levels.

Since the proposed method is an edge based method, noise causes errors and increase the time of caption extraction processing. We perform the noise reduction by low-pass filter [11], so it will remove noise pixels and low intensity edges. The result image is called the filtered gray image '*FG*'.

4.3. Location Extraction

The goal of Location Extraction processing is to extract caption candidate location which may include pixels of caption. Most captions contrast with neighboring pixels regardless of various caption size. Therefore, we extract locations which contain captions from entire images by gray values contrast. However, if contrast information in entire images is used then it may extract background pixels because backgrounds include variety gray values such as same color of captions. Thus, in order to avoid the background locations are extracted from entire image by proposed method, we split *FG* into sub blocks and extract locations which include caption using gray value in blocks.

This step consists of 4 substeps as Figure 12. Proposed method is edge based method. First we extract 8 Edge Direction Images from *FG* in Edge Direction Extraction step. The images are E_1, E_3, E_5, E_6, E_7, E_9, E_{11} and E_{12} where the subscripts are the direction of edges. For example,E_1 contains

only edges of 1:00 direction in *FG*. Ideally pixels of each caption is a uniform color, however, colors of caption pixels in images are not uniform because of image compression. Therefore, we search cross center pixels in Center Pixel Extraction step, but colors of each cross center pixel are not the same. Thus, we extract the representative caption color by clustering colors of cross center pixels, and make Initial Candidate Location Image '*ICL*' which includes pixels of which colors are similar to the representative one. In Location Expansion step, *ICL* is extended by adding neighbor pixels which are similar to colors of pixels in *ICL*, and Caption Location images '*CL*' is produced.

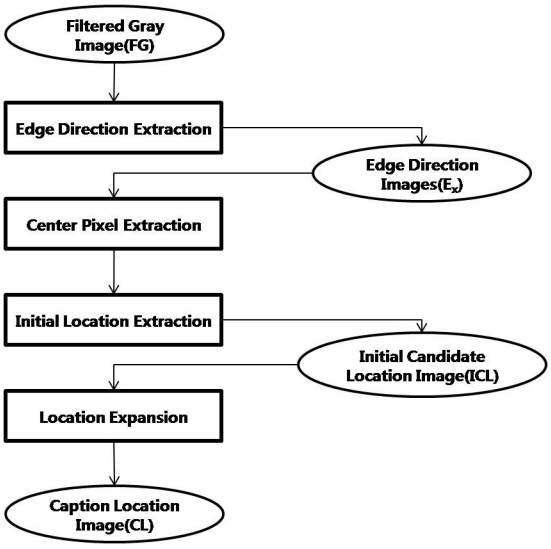


Figure 12. Overview of Location Extraction

Since most captions are contrast to background, edges exist in locations where captions are. Thus, we extract edges to search caption in Edge Direction Extraction step. In order to obtain caption locations, Edge Direction Extraction step extracts all direction edges to make Edge Direction Images 'Ex' using each Roberts crossgradient operators in Eq. (2).

$$\begin{aligned}
E_{11}(x,y) &= FG(x,y) - FG(x-1,y+1)\\
E_{12}(x,y) &= FG(x,y) - FG(x,y+1)\\
E_1(x,y) &= FG(x,y) - FG(x+1,y+1)\\
E_0(x,y) &= FG(x,y) - FG(x-1,y)\\
E_2(x,y) &= FG(x,y) - FG(x+1,y)\\
E_7(x,y) &= FG(x,y) - FG(x-1,y-1)\\
E_6(x,y) &= FG(x,y) - FG(x,y-1)\\
E_5(x,y) &= FG(x,y) - FG(x+1,y-1)
\end{aligned} \tag{2}$$

Since most images are compressed to suit standard format, colors of captions are merged. However, some pixels exist between similar color's pixels. Ideally, colors of each caption are uniform. The center pixels of strokes are less influenced thanother pixels. Therefore, the Center Pixel Extraction step determines pixels which are similar to colors of captions. Since some center pixels locate boundary pixel of caption. Therefore, we search cross center pixels which do not locate boundary of caption using horizontal and vertical cross center pixels because almost cross center pixels locate inner of strokes.

The Initial Location Extraction step extracts pixels which color of pixel is similar to colors of captions. Although cross center pixels are influenced less than other pixels, the colors of captions in entire image are not same color because the each colors of captions are not same others.

In order to extract all caption in images, we split GF into sub blocks. Then representative color of caption in each block is selected by clustering. However, block based method does not guarantee because some block include not enough cross stroke center pixels. Therefore, we make sub-block by shifting it as a half-block.

The input of clustering is gray value of cross stroke center pixels. The clustering process repeats merging two clusters which exist on shortest distance between it. Since caption contrast with background, we decide threshold value of cluster maximum distance. If shortest distance is larger than α, clustering finish (according to experiment, we select that α is 15). At clustering finished, caption representative color decides gray value of cluster which includes highest gray value because gray value of caption is brighter than background.

The gray value of clustering is only one value. Since color of caption is not uniform color, however, the caption representative color in each block is similar to caption's original color. Thus, we classify pixels of block into caption pixels or not by distance of gray value between each pixels gray value and caption representative color. Although color of caption in image is not uniform color by image compression, it is similar to caption's original color. Therefore, pixels which are similar to caption representative color add in Initial Candidate Location Image (ICL).

Some cross stroke center pixels locates boundary of caption because stroke width is not enough to extract cross center pixels such as stroke width shorter than 3. In this case, some cross center pixels locate boundary of caption. In order to solve the problem, Location Expansion step expands caption pixel of ICL to neighbor pixels. The pixels in each block classify into two classes to avoid unnecessary expansion.

First case that gray value of each pixel is brighter than caption representative color in same block. Since caption pixels are brighter than background pixels, pixels of this case add to Candidate Location Image (CL). Second case that gray value of pixel in block is darker than caption representative color in same block. Since pixels of second case contain background pixels. Therefore, second case's pixels need classification that pixel is included by caption or not by caption location expansion mask shown as Figure 13. The mask overlaps CP on caption pixel position of ICL. Almost gray value of cross stroke center pixel is higher than neighbor pixels. Distance of cross stroke center pixels more increase, gray value of pixel more decrease. Therefore, If CP is caption location pixel and x' is non-caption location pixel and x'' is non-caption location pixel (x = 1, 3, 5, 6, 7, 9, 11, 12), we can know x' is boundary of

caption. However, some pixels which satisfy mentioned condition but don't contain caption location. Gray value of x' is lower than gray value of CP and higher than gray value of x''. Therefore, if second case pixels satisfy mentioned condition, the pixels add to Candidate Location Image.

11″		12″		1″
	11′	12′	1′	
9″	9′	CP	3′	3″
	7′	6′	5′	
7″		6″		5″

Figure 13. Caption location expansion mask

4.4. Caption Edge Extraction

In Location Extraction step, we extract locations which contain captions using colors of pixels regardless of various captions size. This step classifies edges of Ex pixels which are located in CL into pixels of captions or not. Figure 14 shows the caption edge extraction processing.

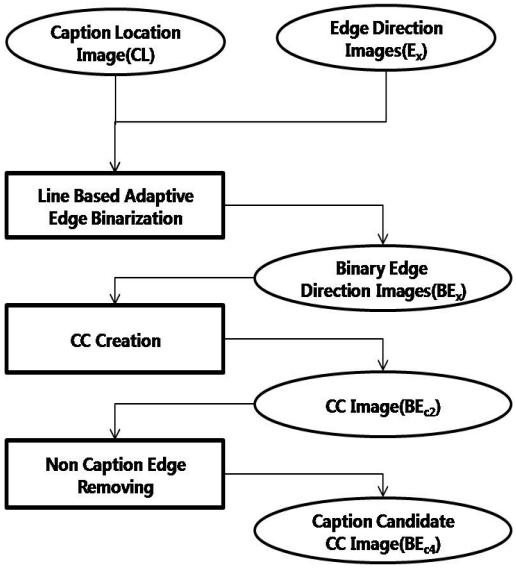


Figure 14. Overview of caption edge extraction step

We extract locations which include edges of captions. However, CL includes edges of captions or not because it are created using only gray values of pixels. And E_{xs} include edges of entire images. In Line Based Adaptive Edge Binarization, first we edges in all Ex remove to check edges in CL for processing time reduction.

We extract CL in Location Extraction step to remove locations where captions do not exist. Most pixels of caption are higher intensity than edges of background. Therefore, we classify pixels in CL into pixels of captions or not in Line Based Adaptive Edge Binarization step. Since some edges have higher intensity than others, we check the corresponding direction edge to remove edges of background. In order to remove edges of background, we create CC to check the corresponding direction edge in CC Creation step. Each edge direction image presents the intensity of the corresponding direction edge. For example, E_{11} image includes 11:00 direction edge's intensity at each pixel position. The other imagescontain the edgeintensity of 12:00, 1:00, 9:00, 3:00, 7:00, 6:00 and 5:00 direction.The edge intensity of caption is usually higher than that of non-caption because values of caption pixel contrast with background gray level values of pixel. However, the intensity difference between captions and background is usually dependent on the size of captions. If the font size is large, the both are significantly contrasted to each other. Otherwise, the boundary between them may have lower intensity edge pixels because small captions include narrow strokes which tend to be much blurred in image compression. So, we remove the pixels of non-caption by an adaptive binarization. We set a threshold for an edge to be a caption edge higher as the font size increases. Then, the question is how to know the font size.

We propose a horizontal line-based edge binarization method which uses 12:00 and 6:00 direction edges. Since Korean syllables include many horizontal strokes, which include vertical, 12:00 or 6:00 direction edges, a feature of a Korean character is the high ratio of 12:00 or 6:00 direction edges. Thus, caption existing area may include more 12:00 or 6:00 direction edges than other area. Especially, the top and the bottom boundary pixels of captions may include more 12:00 and 6:00 direction edges.

First, we define edge pixels of which intensity is greater than as the primitive boundary pixels in E_{12} and E_6. Then we count the number of the primitive boundary pixels in a row. Let us call the number n_T. Then, we count the number of the consecutive primitive boundary pixels in the row, i.e. the primitive boundary pixels of which right or left pixel are also the primitive boundary. Let us call the number n_L. If the value of n_L/n_T is larger than a threshold, we may conclude the row is a top or bottom boundary line of a horizontal sink.

We call such rows the base lines. We set the threshold by analyzing Korean captions, experimentally. The counter, n_L, is the number of pixels which possibly belong to a horizontal stroke, because its adjacent pixel is also an edge pixel. So the higher n_L/n_T of a row is, the more probably the row is the top or the bottom line. For example, let us suppose that Figure 15 is an image of 12:00 and 6:00 edges of 3 Korean syllables. For simplicity, 12:00 and 6:00 edges are presented in white, other direction edges in gray and background in black. Table 1 shows n_L, n_T and n_L/n_T of each row of Figure 15. The top line of the captions in Figure 15 is the second row and the bottom boundary is the 18th row. From Table 1, we may know that the ratio of n_L/n_T of the both is higher than other rows. If we carefully select the threshold for the base line, then we can mostly find the top and the bottom row. So, the distance between the baselines may be regarded as the font size.

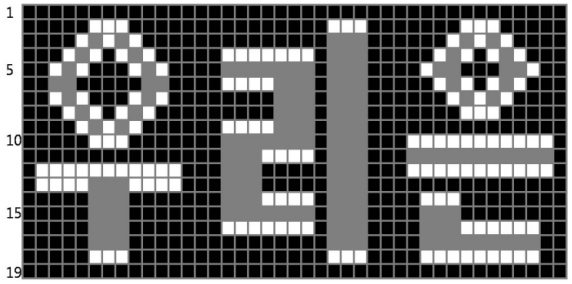


Figure 15. Distance image of 12:00 direction edge

So, we set the threshold of gray level for detecting edge pixels proportionally to the distance between neighboring base lines as equation (3).

$$t(y) = d(y) \times \beta + \gamma \tag{3}$$

In Eq. (3), $t(y)$ is the threshold value of the yth row. $d(y)$ is the distance between the upper and the lower base lines of the yth row. and are constants. Then, we apply the threshold to all the yth rows of E_{xs} and get the binarized image BE_{xs}.

Each BE_x may include edge pixel of captions as well as background. However, the pixels of captions always have the corresponding edge. We can remove the pixels of background by checking corresponding edges. To check corresponding edges, we make CC images from BE_{xs}.

Table 1. Ratio ofconsecutive primitive boundary to primitive boundary of each lines

row	n_L	n_r	n_r/n_L
1	0	0	0.00
2	9	9	1.00
3	13	0	0.00
4	22	7	0.32
5	22	0	0.00
6	20	4	0.20
7	17	0	0.00
8	15	3	0.20
9	15	4	0.27
10	24	14	0.58
11	21	4	0.19
12	28	25	0.89
13	17	8	0.47
14	16	7	0.48
15	16	0	0.00
16	22	13	0.59
17	15	0	0.00
18	15	15	1.00
19	0	0	0.00

First, we make BE_{all} by merging all BE_{xs}, and then create the CC image, BE_{c2}, by identifying connected components in BE_{all}. Then, we search the corresponding edges for all edges in BE_{c2}. If the corresponding edge of an edge is not found, the edge pixel is removed from BE_{c2}. Finally we again build CCs from the remaining pixels in BE_{c2} and get BE_{c4}, the binarized image with caption candidate CC. Even though we removed the edges without the corresponding edge, BE_{c4} may include many non-caption edges because non-caption edges may also have the corresponding edge. So, we need more strict classification step.

4.5. Caption CC Extraction

The goal of this step is to discriminate caption CCs from non-caption CCs. This step consists of 3 sub steps as shown Figure 16. The first step is Edge Refinement. Edge Refinement step extracts the strict boundary edges from BE_{xs} by applying refinement makes. CC Classification step classifies the CCs into caption CCs or non-caption CCs by a decision tree which was made in advance. However, the classification result may have some error, thus we correct the classification result in Caption Restoring step. The purpose of Edge Refinement is to prepare the features of CCs which will be the inputs of the decision tree. A decision tree will discriminate caption CCs from non-caption CCs with the features.

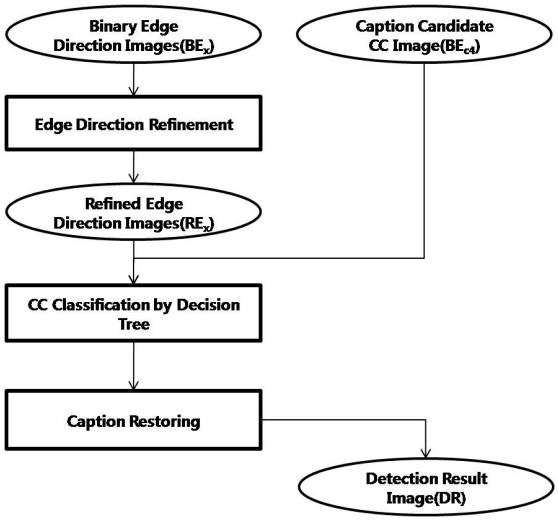
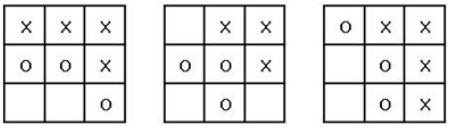
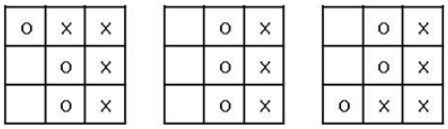
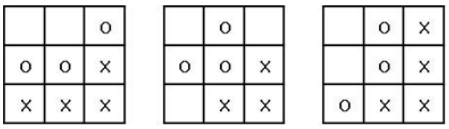
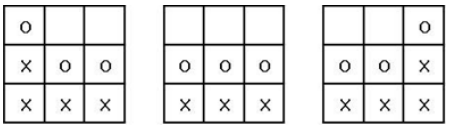
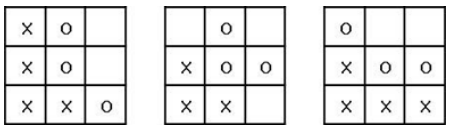
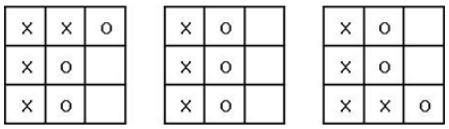
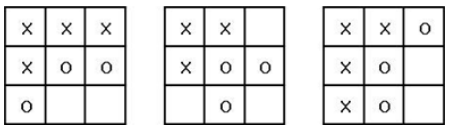
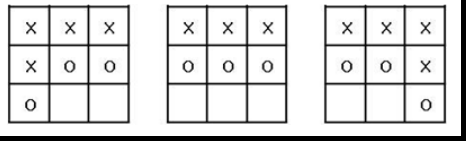


Figure16. Overview of caption CC extraction

The input of Edge Refinement step is BE_{xs}, binarized edge images, and the output is refined binarized edge images, RE_{xs}. When we create E_{xs}, which is the original edge image of BE_{xs},

we use Roberts's masks to obtain all potential caption edges. So there are many edges not only from captions but also from background in BE_{xs}. In order to extract more strictly caption edges from BE_{xs}, we apply the masks in Table 2. We apply three masks to each image of BE_{xs}. For example, BE_{11} includes 11:00 direction edges, so we specially design the masks to strictly extract 11:00 direction edges as shown in the 7th row of Table 2.

Table 2. Edge refining masks

Target image	Masks
BE_1	
BE_3	
BE_5	
BE_6	
BE_7	
BE_9	
BE_{11}	
BE_{12}	

The center of each mask locates at a pixel. If the 'O's correspond to an edge pixel and the 'X's do not, then the pixel is checked as a strict caption edge pixel. This mask processing makes the refined edge images: RE_1, RE_3, RE_5, RE_6, RE_7, RE_9, RE_{11} and RE_{12}.

Using RE_xs and BE_{c4}, we generate the features of caption CCs. The features are the ratio of each 1:00, 3:00, 5:00, 6:00, 7:00, 9:00, 11:00 and 12:00 direction strict edges. In order to check if a CC in BE_{c4} is a caption CC or not, we count all edges in the CC in BE_{c4}, let us call the number E, and all edges which locate in BE_x corresponding to the locate of the CC in BE_{c4}, let us call number E_x. Then E_x/E is the ratio of x direction strict edges. In this way, we evaluate the direction edge ratios. Those edge ratios are useful for caption CC discrimination, because captions mostly include certain direction strokes. The distribution of edges in caption CCs are different from that in non-caption CCs. There are two more features: the standard deviation of center pixel's intensity between horizontally corresponding edges (HSD) and the standard deviation of center pixel's intensity between vertically corresponding edges (VSD). Those features are to check the color between edges. If a CC is a caption CC, then the color of pixels between horizontal or vertical edges are very similar to each other, so the deviation may be small. Those 10 features are the input of the decision tree.

Our decision tree uses ratios of each x direction strict edges, HSD and VSD, and was built using regression tree algorithm. Figure 17 is the decision tree which is using 10 features and is limited to tree depth of 5 levels for readers understanding easily. Each internal node, which draws circle in Figure 17, is the condition of classification. And each terminal node, which draws rectangle in Figure 17, is result of classification.

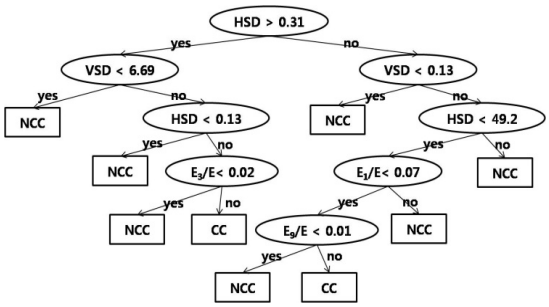


Figure 17. Created decision tree

The decision tree classifies CCs in BE_{c4} into caption CCs or non-caption CCs. If a CC is classified as a non-caption CC, then it is removed. Since decision tree algorithm always does not guarantee correct classification, the result of discrimination may not perfect. Especially, in small caption CCs, strict edges may not be correctly extracted because there are many short and thin strokes. Thus it is probable that small caption CCs are removed by the decision tree. So, we add Caption CC Restoring Step.

Restoring is based on an assumption that neighboring characters in a row will appear at the same horizontal level with the same height, and at least three characters will appear in a row. Restoring is done in two steps: caption CC restoring and non-caption CC discarding. First, caption CC restoring is to restore caption CCs removed by the decision tree. We simply restore a removed CC if its vertical position and height are almost same as those of its neighboring CCs which are

classified as caption. Based on our assumption, such a CC is strongly regarded as a caption CC. Next, in order to discard non-caption CCs among CCs classified or restored as caption, we count the number of CCs at the same vertical position of a CC. If there are less than 3 CCs, the CC is discarded because we assume that at least three characters will appear in a row. Finally we have highly probable caption CCs in *DR*.

4.6. Caption Edge Extraction

The goal of this step is to separate the caption pixels from the gray-converted original image *G*. The output of this step is a binarized character image which can be inputted into an OCR engine. However, If stroke width of caption is wide such as wider than 8 pixel size then all pixels of caption may not contain edges because the neighbor pixels of cross center are not influenced by image compression. Therefore, we convert *G* into text image using *DR* and *G*.

For separation, we have to identify the position of characters and the threshold level for binarization. The final CCs in *DR* are the connected components of caption edges. So the position of a character is simply obtained from the position of a CC. Next, to obtain the binarized character image where background pixels are converted into black and character pixels are into white, we evaluate the average intensity of pixels in *G* corresponding to the edge pixels in RB_{xs}. RB_{xs} include the strict edges which are the highly probable boundary of background and character pixels, so in image *G*, the intensity of the inner pixels of captions will be higher than that of boundary pixels. We convert the pixels with higher intensity than the average into white and vice versa.

5. Experiments

The proposed method aims to extract captions which are various font types and sizes from video images. It can also reduce some unnecessary processes which are included in the existing methods. There are some goals as follows:

- Extracting the latent area of captions (Location extraction)
- Extracting the connected objects which contain captions (Caption edge extraction and Caption CC extraction)
- Generating the caption images which are binarized images of captions and background using a binarization of the found connected objects (Text segmentation)

We examine our proposed method at each phrase for verifying it. Generally, captions are related with its images except for meaningless words such as onomatopoeia and mimetic word. So, we collect the video clips of documentaries, news and sports videos which contain meaningful captions for the contents. There are total 168 video clips for the test, and they have 4,729 Korean characters in the captions. There also exist also various font sizes from 8 to 60-fixel.

The proposed method reduces candidate captions gradually, that is based on the eliminatingdefinite non-caption areas. The first step of the eliminating non-caption areas is extracting the areas which have captions potentially. This process is done at the location extraction step using the distribution information of cross center pixels.

The image data collected from broadcast programs such as documentaries, news programs, and sports programs that caption play an important role to information transmission. The gathered image includes caption with various fonts and sizes which include 12 ~ 45 pixel-height.Existing methods proposed evaluation technique for caption extraction methods compare detected caption location and actual caption location, or compare extracted CC and actual CC, etc. however, the existed methods consider that caption extraction evaluation is important in detecting all captions, as well as in detecting only actual captions. Therefore, recall and precision criteria were used. Recall is the ratio of the correctly identified captions to all the captions in an image, while precision is the ratio of the correctly identified captions to extracted captions.

Table 3 shows existed method and proposed methods experimental result. Proposed methods recall rate is lower than Precision because our methods can extracts caption which are scene text as score board or sponsor banner in image by CC's pattern.

Table 3. Experimental results

Methods	Recall	Precision	Data Type	Object Language
[2]	94.8%	95.3%	Image	-
[8]	88.0%	-	Video	-
[12]-1	91.7%	86.0%	Image	Chinese
[12]-2	84.5%	74.0%	Image	Chinese + English
Proposed method	**88.3%**	**78.8%**	**Image**	**Korean**

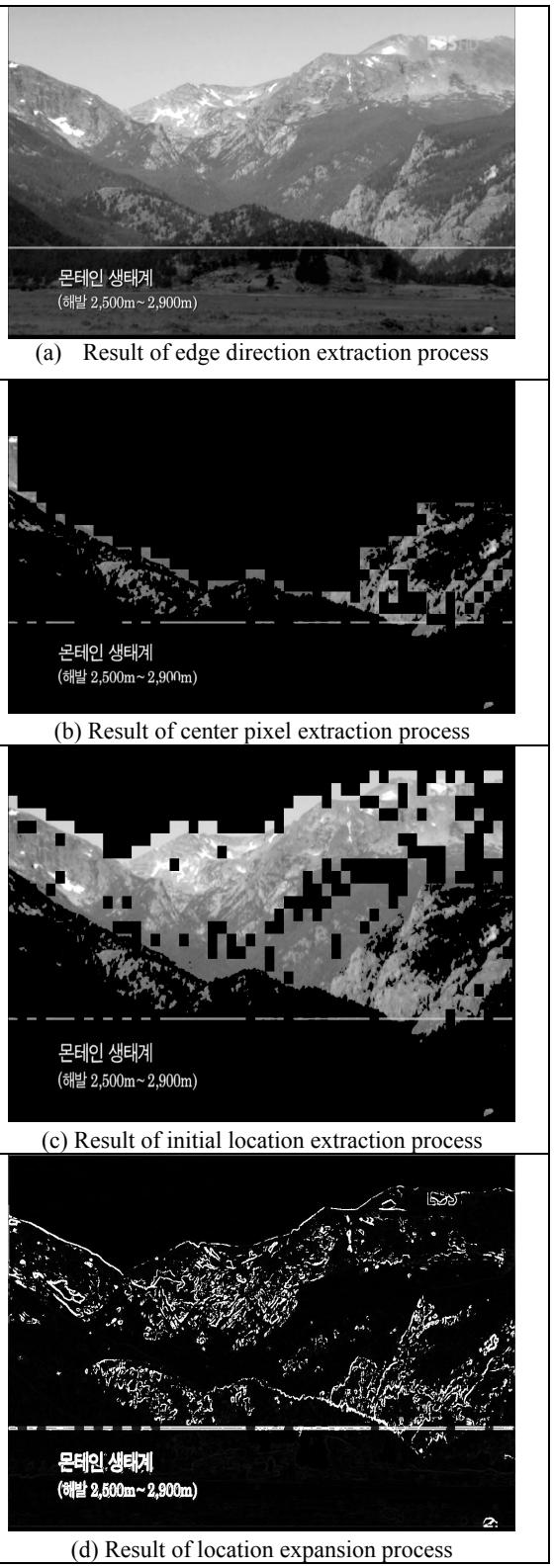
The existed methods in Table 3 propose language independent methods, they not scribe extraction rate by captions sizes. Also, they use their collected data, direct comparing is impossible. Therefore, we show extraction rate in Table 4.

Table 4. Result of various size caption extraction

Height (pixel-size)	# of caption	# ofextracted caption	Recall
~ 10	308	265	86.0%
10~20	2,071	1,830	88.3%
20~30	918	813	88.5%
30~	1,504	1,334	88.6%

We decide all caption size as pixel-height and evaluate caption extraction performance by caption sizes. However, precision can't be evaluated by us because non-caption CCs exists various size. Number of caption which its size shorter than 10 pixel-height are smaller than other size caption.

Figure 18(a) ~ (d) shows the result of each process of the proposed method.



(a) Result of edge direction extraction process

(b) Result of center pixel extraction process

(c) Result of initial location extraction process

(d) Result of location expansion process

Figure 18. The result of each process of the proposed method.

As we mentioned at the subsection 4.3, after edge direction extraction, we can get 8-direction information from images. At this step, the colors of caption pixels in images are not uniform because of image compression. So we apply the center pixel extraction process as Figure 18(b) shows. This step helps us to identify the colors. And we can also consider the difference of each color of cross center pixels by applying the Initial location extraction in Figure 18(c). Finally, we can extract the caption location images after using location expansion step as Figure 18(d) shows.

6. Conclusion

We propose a method which can extract Korean captions from images which contain captions of various sizes and fonts by the decision tree algorithm. Existing extraction methods extract captions of various sizes by the sequential multi-resolution method that extracts captions repeatedly to modify resolution images. And most existing methods also consider only language dependent morphological features. However, the proposed method can extract captions without those processes and limitations. Further research will be focused on two objectives. First, the proposed method is expanded to sequential images for improvement of precision and recall. And then the connected NCCC and CCC spilt up each CC, or the noise reduction processing alternative another algorithm. The proposed method extracts edges simply by an edge extraction method and removes the noise edge which is low intensity by an adaptive line-based edge threshold method. Then, the remaining edges which are not included in CCC are removed by checking the corresponding edge. And the decision tree algorithm classifies CC into CCC and NCCC.

References

[1] R. Lyu, J. Song, and M. Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization and Extraction,"*IEEE Transaction on Circuits and Systems for Video Technology*, vol. 15, no. 2, pp. 243-255,2005.

[2] J.-M. Jeong, J. Cha, and K. Kim, "A Stroke-Based Text Extraction Algorithm for Digital Videos," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 17, no. 3, pp. 297-303, 2007.

[3] K.C. Jung, K.I. Kim, and A.K. Jain, "Text Information Extraction in Images and Video: A Survey,"*Journal on Pattern Recognition*, vol. 37, no. 5, pp. 977-997, 2004.

[4] E.K. Wong and M. Chen, "A Robust Algorithm for Text Extraction in Color Video," *the Proc. of the IEEE Multimedia and Expo 2000 (ICME 2000)*, vol. 2, pp. 797-800, 2000.

[5] K.C. Jung and E.Y, Kim, "Automatic Text Extraction for Content-Based Image Indexing,"*Lecture Notes in Computer*

Science, vol. 3056, pp. 497-507, 2004.

[6] J.-H. Jung, T.-B.Yoon, D.-M.Kim, and J.-H.Lee, "Connected Component-Based and Size-Independent Caption Extraction with Neural Networks,"*Journal of Korean Institute of Intelligent Systems*, vol.17, no.7, pp.924-929, 2007.

[7]J.-M.Jeong, J.-H.Cha, and K.-H. Kim, "A Stroke-Based TextExtraction Algorithm for Digital Videos,"*Journal of Korean Institute of Intelligent Systems*, vol.17, no.3, pp.297-303, 2007.

[8]H. Byun, I. Jang, and Y. Choi, "Text Extraction in Digital News Video Using Morphology," *Lecture Notes in Computer Science*, vol. 2423, pp. 341-352, 2002.

[9]Y.M.Y. Hasan and L.J. Karam, "Morphological Text Extraction from Images," *IEEE Transactions on Image Processing*, vol. 9, no. 11, pp. 1978-1983, 2000.

[10]H.E. Jiaying, L.I. Shaofa, "Hybrid Chinese/English Text Identification in Web Images," *Proc. of the 3rd International Conference Image and Graphics (ICIG '04)*, pp. 361-364, 2004.

[11] www.Stephen Wright.org/Korean

[12] J. Song, M. Cai, and M.R. Lyu, "A Robust Statistic Method for Classifying Color Polarity of Video Text," *Proc. of the IEEE International Conference Acoustics Speech and Signal Processing(ICASSP '03)*, vol. 3, pp. 581-584, 2003.

Je-Hee Jung
M.S. Student of Sungkyunkwan University, Korea
Research Area: Image processing, Intelligent system, Etc.
E-mail: gullingi@skku.edu

Jaekwang Kim
Ph.D. Candidate Student of Sungkyunkwan University, Korea
Research Area: Web data mining, Intelligent system, Etc.
E-mail: linux@ece.skku.ac.kr

Jee-Hyong Lee
Professor of Sungkyunkwan University, Korea
Research Area: Intelligent system, Fuzzy theory
E-mail: jhlee@ece.skku.ac.kr