J Real-Time Image Proc (2016) 11:167–177
DOI 10.1007/s11554-012-0298-x

ORIGINAL RESEARCH PAPER

Korean-English bilingual videotext recognition for news headline generation based on a split-merge strategy

Cheolkon Jung · Licheng Jiao

Received: 25 October 2011 / Accepted: 25 October 2012 / Published online: 17 November 2012

Abstract This paper deals with Korean-English bilingual videotext recognition for news headline generation. Because videotext contains semantic content information, it can be effectively used for understanding videos. Despite its usefulness, it is a challengeable task to apply text recognition technologies to practical video applications because of the computational complexity and recognition accuracy. In this paper, we propose a novel Korean-English bilingual videotext recognition method to overcome the computational complexity as well as achieve comparable recognition accuracy. To recognize both Korean and English characters effectively, the proposed method employs an elaborate split-merge strategy in which the split segments are merged into characters using the recognition scores. Moreover, it avoids unnecessary computation using geometric features such as squareness and internal gap, and thus its computational overhead is remarkably reduced. Therefore, the proposed method is successfully employed in generating news headlines. The effectiveness and efficiency of the proposed method are verified by extensive experiments on a challenging database containing 51,290 text images (176,884 characters).

Keywords Korean-English videotext recognition · Content-based video retrieval · News headline generation · Split-merge strategy · Video OCR

C. Jung (✉) · L. Jiao
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China
e-mail: zhengzk@xidian.edu.cn

1 Introduction

With recent advances in digital broadcasting and internet technologies, smart TV which integrates the internet into digital TV set-top boxes has become increasingly popular. It allows viewers to search and find digital videos, photos, and other contents on the web or stored on a local hard drive. Thus, it is very important to analyze and index such media for better understanding of contents. The related research issues are automatic annotation, indexing, summarization, and retrieval of digital videos. The chief goal of the research is to index video data efficiently, make it a well-structured media form, and provide viewers with intelligent content-based browsing and retrieval functions.

During the past decade, a lot of great results in content-based video indexing and retrieval have been achieved by researchers [1–20]. Dimitrova et al. [3] introduced key technologies and applications for video-content analysis and retrieval. They addressed basic technologies for content-based video management based on user's needs. Above all, they introduced several enterprise and consumer domain applications. The enterprise applications were professional and educational ones made by broadcasting companies and content providers. The consumer domain applications were mainly related to the content-based video management technologies developed at Philips Electronics such as Vitamin and Video Scout [4, 5]. Kim et al. [6] presented a news video summarization method based on multi-modal analysis of contents. The proposed method employed closed caption (CC) data and speech signals in audio stream. The speech signals were utilized to align and synchronize the CC data with the video in a time line. In addition, semantic highlights were created by the CC data and described in a multilevel structure using the MPEG-7 summarization description scheme (DS). They also introduced the summary generator

and video browser which could retrieve video clips by inputting text query. Merialdo et al. [7] presented an automatic construction method of personalized TV news programs, which could provide predefined duration and maximum content value to a specific user. In this method, video indexing and information filtering techniques were efficiently combined to select stories which were most adequate given the user profile. Their contribution was to create the customized program with a predefined duration that had been specified by each user based on user preference.

Liu et al. [8] presented an advanced content-based news video browsing and retrieval system (NewsBR), which provided users with high-accuracy news story segmentation and topic generation of each news item. For the high accuracy story segmentation of news videos, silence clips were detected and successfully combined with shot boundary detection results. Thus, the NewsBR system was helpful for users to understand news videos quickly.

In this study, we also aim to index and annotate broadcast news videos to provide users with intelligent services. Because users generally want to preview the main topics of news items and quickly decide what they want to view in more detail, headline generation in news videos is required. Videotexts are very useful for generating news headlines because they contain semantic information and summarize each news item concisely. However, videotext images often contain overlapping and touching characters as shown in Fig. 1. As can be seen, touching parts of text images appear in the red dotted circles. The touching parts have bad effects on the recognition rate. In general, they are mainly caused by the character degradation and stroke distortion from lossy video compression. Thus, accurate character separation is required for successfully recognizing videotexts.

Videotext recognition, i.e., video optical character recognition (OCR), was first introduced by the Sato et al. [16, 17]. In the methods, a simple pattern-matching technique was used to recognize videotexts. Text images were normalized in size and converted into a blurred gray-scale image to make the recognition robust to changes in thickness and position. Then, the normalized gray-scale images were matched and identified with reference patterns using a correlation metric. Chang et al. proposed a prototype classification method for efficient training of support vector machine (SVM) [18]. Because SVM was extremely slow in the training process when the number of target classes was large, the candidate classes for SVM were selected by

Fig. 1 Examples of touching characters in videotext images. *Left* English, *right* Korean. Touching parts appear in the *red dotted circles*

K-means clustering in the proposed prototype classification method. Thus, this method reduced the number of binary classification problems and made it possible to use SVM for large-scale character recognition. Lee and Kim [19] proposed a complementary combination method of two recognition methods: a holistic-based recognition and a component-based recognition. The holistic-based recognition method employed the global shape information of a character image, while the component-based recognition method used a detailed local shape of each text segment. In this method, the two recognition methods were elaborately combined to improve the recognition accuracy of videotexts. Park et al. [21] proposed a recognition method for Korean characters in outdoor scenes. This method utilized a minimum distance classifier with a shape-based statistical feature for character recognition.

Most work on videotext recognition utilized a commercial OCR to perform recognition [20–30]. Thus, they were mainly focused on improving text segmentation, i.e., two-level thresholding, before the recognition by an OCR module. Even if many commercial OCR systems worked well on good-quality scanned documents under controlled conditions, there were limits to the recognition of videotexts. The commercial OCR systems were usually trained by text images with high quality, however, most videotext images had low resolution because of the loss of high-frequency components caused by lossy video compression [19, 22]. Actually, videotext images had different properties from the characters of scanned documents including large stroke distortion, font variation, and variable size. In addition, in the case of multilingual texts, the computational complexity is dramatically increased because the number of target classes becomes very large. Therefore, we need an effective multilingual videotext recognition method considering the computational complexity and recognition accuracy. In this paper, we propose a novel Korean-English bilingual videotext recognition method to improve both the computational complexity and recognition accuracy. The overall process of the proposed method is illustrated in the block diagram of Fig. 2. As shown in the figure, the proposed method is comprised of four main modules: special-character filtering, character separation, character recognition, and character refinement. The proposed method employs an efficient split-merge strategy which is very effective in separating and recognizing the text images of news videos as well as possesses the capability of recognizing both Korean and English characters. It avoids unnecessary computation using geometric features such as squareness and internal gap, and thus its computational overhead is remarkably reduced.

The rest of this paper is organized as follows. Section 2 describes the proposed videotext recognition method in detail. In Sect. 3, experimental results are shown. Finally, conclusions are made in Sect. 4.

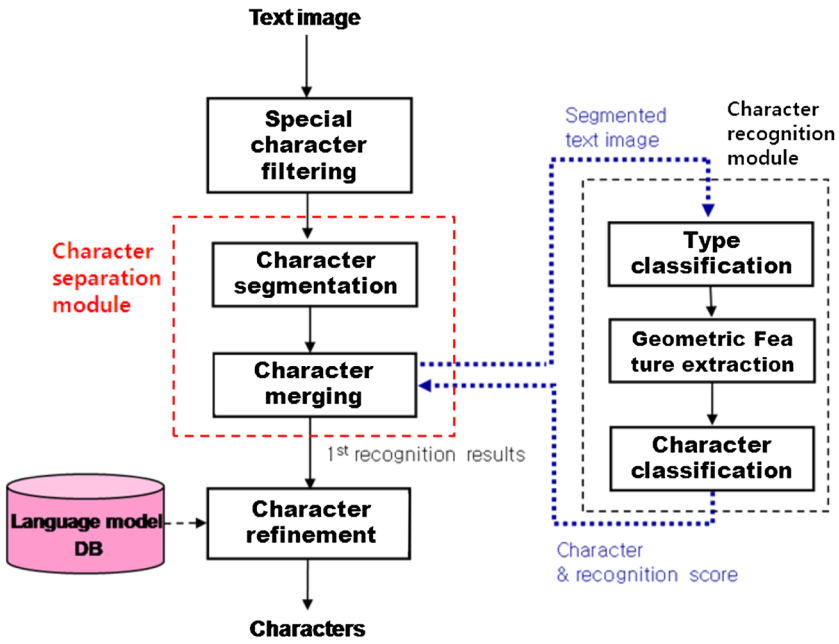

Fig. 2 The overall process of the proposed videotext recognition method. The proposed method is comprised of four main processes: special-character filtering, character segmentation, character merging, and character refinement

2 Proposed method

The proposed videotext recognition method is comprised of four main modules: special-character filtering, character separation, character recognition, and character refinement. The character separation module consists of character segmentation and merging processes and uses the feedback from the recognition score of the character recognition module.

2.1 Special-character filtering

Before recognizing videotexts, predefined special characters are filtered using heuristic information from text images. The predefined special characters are seven classes of ( ) “ ‘ ° . ,. The seven classes often appear in news and sports videos and have detrimental effects on the recognition rate. Thus, they should be handled before recognition. The five classes of “ ‘ ° . , are classified by position information as shown in Fig. 3. If the position of segments is above the middle line in Fig. 4, the segments are classified as one of the three special-character classes: “ ‘ °. In addition, if the position is below the middle line, the segments are classified as one of the two classes: . ,. Then, the rest two special-character classes of ( ) which are parentheses, are classified using template matching. The templates of the parentheses are generated by overlapping training images into a gray-scale image as shown in Fig. 4.

2.2 Character separation

The next step is the character separation module which is a combination of character segmentation and merging

Fig. 3 Handling special characters by position information

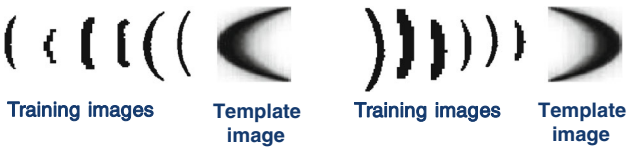

Fig. 4 Handling special characters by template matching

(see Fig. 2). The proposed method employs a split-merge strategy in which the split segments are merged into characters using the recognition score obtained by the character recognition module [15, 31]. In general, there are several characters in input videotext images, and thus the character separation module that decomposes the text image into sub-images of individual characters is required. The character separation module is very important because the character separation performance significantly affects the accuracy of character recognition. However, it is not easy to separate characters accurately because most of videotext images contain overlapping and touching parts between two characters (see Fig. 1). Therefore, our character separation module uses the feedback from the recognition score of the character recognition module. Thus, when we finish the task of the character separation module, both recognition and separation results are obtained. The separation module employs the split-merge strategy in which the split segments are merged into a character. Thus, the pre-segmentation stage to segment the

input text image into several segments is needed as shown in Fig. 5. In the proposed method, the nonlinear cutting path method proposed by Tseng et al. [32] is used to effectively segment the overlapping and touching characters. In this method, character string is regarded as a multi-stage directed graph for generating nonlinear cutting paths in videotext images. Observation score for each pixel and transition score from the top pixels to the bottom pixels are defined in advance. The less number of black pixels a segmentation path passes and the more straight it is, the higher score it gets.

Thus, possible paths from the top end to the bottom end of the character string are selected using the Viterbi algorithm. As shown in the figure, the nonlinear cutting path segments overlapping parts and touching regions effectively.

2.3 Character recognition

The proposed method possesses the capability of recognizing both Korean and English characters. Korean characters are structurally more complex than English characters, and the number of target classes in Korean is incredibly more than that in English. Thus, the number of target classes should be reduced for accurate recognition, and thus type classification is required. As a result, our character recognition module is composed of type classification, geometric-feature extraction, and character classification (see Fig. 2). The goal of the type classification is to solve the complexity problem caused by the number of target classes. In our method, characters are divided into small groups based on characteristics of characters. As shown in Fig. 6, characters are classified into seven types: six types of Korean characters and one type of alphanumeric- and special-characters [33, 34]. In the figure, FC is the first consonant; VV is the vertical vowel; HV is the horizontal vowel; and LC is the last consonant. To classify the types of characters, the SVM classifier for multi-class problem is employed. Thus, the seven SVM machines are utilized to classify the types of characters and we choose the class which has the maximum output from the seven SVM machines. In the SVM classification, we use the angular directional feature (ADF) as the feature for classification [35]. ADF is obtained by calculating the number of eight angular directions in a mesh window as shown in Fig. 7.

Fig. 5 Pre-segmentation results generated by nonlinear cutting paths

As shown in the figure, feature dimension of the ADF is total 288, i.e., 6 rows × 6 columns × 8 directions, and angular direction α on each pixel (x, y) is determined by:

$$\alpha(x,y) = \tan^{-1}\left(\sum_{a,b} a \cdot f(x+a,y+b), \sum_{a,b} b \cdot f(x+a,y+b)\right), \quad -\frac{w-1}{2} \le a,b \le \frac{w-1}{2} \tag{1}$$

where w is a small window; and a, b are x- and y- indexes in w, respectively. The extracted ADF is also used for character recognition. After classifying the type of characters, geometric-feature extraction and character classification procedures are sequentially performed. The optimal segmentation path is determined by the two procedures. The two scores are calculated to determine the optimal segmentation path from the two procedures. They are the geometric score, S_G, to estimate the likelihood of being a character by geometric features and the recognition score, S_R, obtained by the character recognition process.

First, S_G is computed using the two character evaluators, the squareness (SQU) and the internal gap (GAP). The SQU is determined by the following equation:

$$\mathrm{SQU} = \frac{\mathrm{Min}(\mathrm{CW}, \mathrm{CH})}{\mathrm{Max}(\mathrm{CW}, \mathrm{CH})} \tag{2}$$

where CW and CH are the width and height of each segment. The probability density function (PDF) of SQU, P(SQU), is estimated from the videotext images using Parzen windows [36]. The estimated PDFs of Korean and English characters are shown in Fig. 8. The GAP means the distance between two neighboring segments, and its distribution, P(GAP), is also estimated by Parzen windows. Then, S_G is computed as follows:

$$S_G = \lambda_{\mathrm{SQU}} \cdot \log(P(\mathrm{SQU})) + \lambda_{\mathrm{GAP}} \cdot \log(P(\mathrm{GAP})) \tag{3}$$

where λ_{SQU} and λ_{GAP} are weighting constants. If S_G is low, the segment is eliminated in the paths and not recognized. Thus, the computational overhead of the recognition is reduced.

Next, S_R is determined by computing the distance between the recognition model and each segment. Based on the extracted ADF, S_R of each segment is obtained by the linear discriminant analysis (LDA) [31, 37]. LDA is a

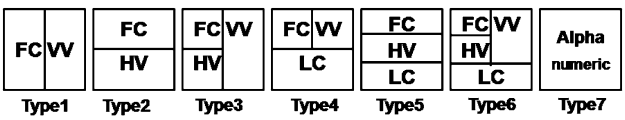

Fig. 6 The seven types of characters: six types of Korean characters (*Type 1–6*) and 1 type of alphanumeric- and special-characters (*Type 7*). *FC* first consonant, *VV* vertical vowel, *HV* horizontal vowel, *LC* last consonant

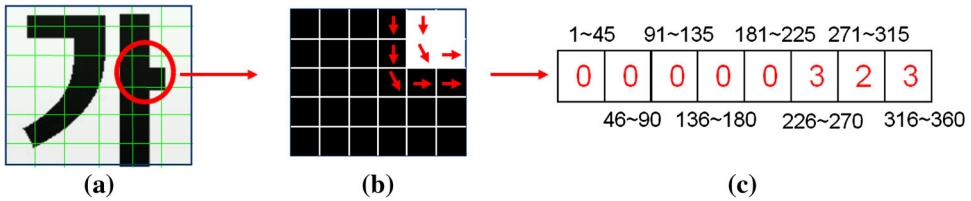

Fig. 7 Illustration of the ADF: **a** Mesh image. **b** One mesh (*red arrows* are directions of pixels). **c** Distribution of angular directions (*black numbers* degree of eight directions, *red numbers* the number of each angular direction)

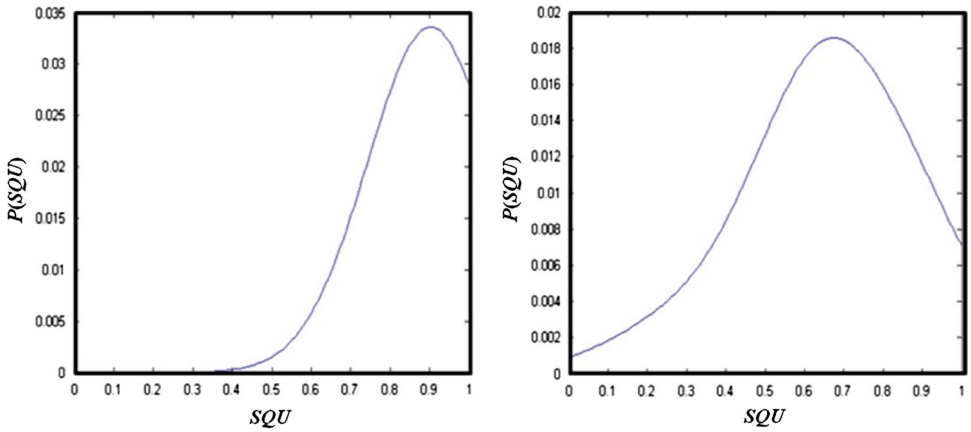

Fig. 8 The estimated PDFs of squareness (SQU) by Parzen windows. *Left* Korean characters, *right* English characters

classification method based on the Mahalanobis distance, and assumes that covariance of each class is the same. Thus, S_R is obtained by the following equation [31]:

$$S_R = 2\mu_i^T \Sigma^{-1} \mathbf{x} - \mu_i^T \Sigma^{-1} \mu_i + 2 \log P(c_i) \quad (4)$$

where $\mathbf{x}$ is a feature vector; c_i is the ith class; μ_i is the mean of the ith class; Σ is the covariance of the class; and $P(c_i)$ is the prior probability of c_i. The average recognition score S_A of each segment is represented by these two scores:

$$S_A = \frac{1}{N} \sum_{i=1}^{N} S_i \quad (5)$$

$$S_i = k_G \times S_G + k_R \times S_R \quad (6)$$

where i and N denote the index and the number of characters; k_G and k_R are constants; and S_i, S_G and S_R denote the total, geometric, and recognition scores of each character, respectively. Thus, the optimal segmentation path is generated in the character merging process based on the average recognition score S_A as shown in Fig. 9. In the figure, the red line is the optimal segmentation path and the blue dotted line is eliminated paths because of the low geometric scores.

2.4 Character refinement

We get the labels and scores of the top ten recognition candidates on the merged results after the character merging process. Also, we obtain the language model scores of the top ten recognition candidates using a language model. The language model scores are combined with the recognition scores on the top ten recognition candidates to obtain the final recognition results, R', as follows:

$$R' = \arg \underset{R}{\text{Max}} [\log P(\mathbf{X}|R) + \log P(R)] \quad (7)$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l]$ is the feature vector of the merged results; R is the recognized character string by the character recognition module; $\log P(\mathbf{X}|R)$ is the average recognition score of each character; and $\log P(R)$ is the language model score. As shown in Fig. 10, we use the word-based language model according to the language type because the text string contains both Korean and English words [38, 39].

3 Experimental results

To verify the effectiveness and efficiency of the proposed method, extensive experiments are performed on a PC with a Pentium IV 2.4 GHz and 2.00 GB memory using Visual C++ 6.0 based on the Window XP operating system.

3.1 Construction of training and testing database

The ground truth database consists of news and sports videos during the 3 months broadcasted in South Korea in 2005.

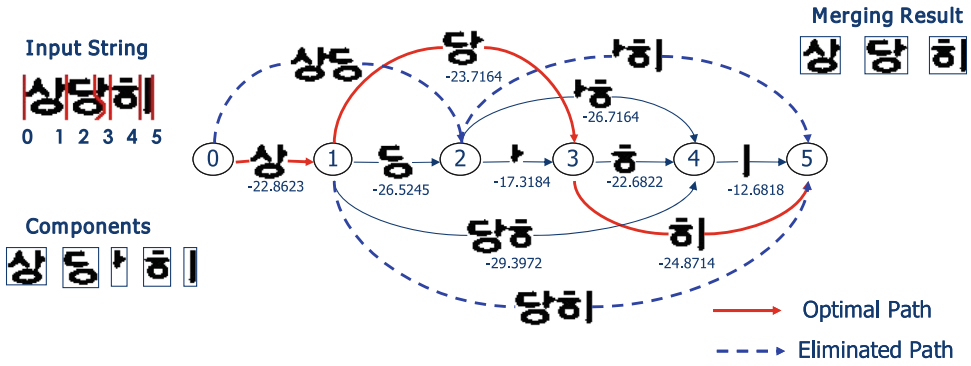

Fig. 9 Example of finding the optimal segmentation path

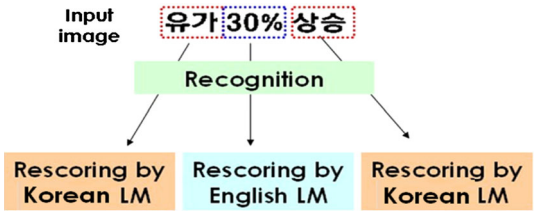

Fig. 10 Character refinement using the language models of Korean and English. *LM* language model

The channels were KBS, MBC, SBS, CNN, BBC, and NBC. Our experimental database contains 51,290 text images (i.e., 176,884 characters) extracted from the news and sports videos, and thus the database comprises a variety of cases, including various texts with a wide spectrum of font-sizes and poor quality, etc. In the database, 36,430 text images (i.e., 125,636 characters) are for training and 14,860 text images (i.e., 51,248 characters) for testing.

In the testing data sets, we get 14,803 text images (i.e., 50,894 characters) from standard-definition television (SDTV, 640 × 480) videos and 57 text images (i.e., 354 characters) from high-definition television (HDTV, 1,920 × 1,080) videos. We use a tool named the *Binarizer* implemented by ourselves to manually or semi-automatically collect the ground truth character boxes and their attributes as shown in Fig. 11. The details of the file structure for the ground truth database are given below. The record of the file consists of seven fields: Label, ImageID, StartX, StartY, EndX, EndY, and Order. Here, the ImageID field stores the file name of the text image; the StartX and StartY store x- and y- coordinates of the start point, respectively; the EndX and EndY store x- and y- coordinates of the end point, respectively; and the Order stores the order of each character in a string. An example of the file structure in the ground truth database is shown in Table 1. In the *Binarizer*, there are two ways to get the ground truth database. The first way is to create the character boxes and their attributes using a fully manual method in the bottom left window of Fig. 11. In this way, we draw a bounding box for each character in the clip by dragging the mouse over the character area, and then enter the information of each character such as label and order. The second way is to generate the character boxes and their attributes using a semi-automated method in the bottom left window of Fig. 11. In this way, characters are automatically separated from text images using the simple projection profile analysis [40]. Thus, we only enter the label field of each character for the ground truth database.

3.2 Performance evaluation of each module

Before evaluating the proposed method, we provide some character separation and recognition results in Fig. 12. As shown in the figure, there are two rows in each result: the first row shows a text image and its separation results while the second one shows the recognition result corresponding to the text image. As can be seen, videotext images even with overlapping and touching characters are successfully decomposed into characters by the proposed method. Moreover, the proposed method is very effective in recognizing videotext images with both Korean and English characters.

In our method, there are four main modules for videotext recognition: special-character filtering, character separation, character recognition, and character refinement. Because the special-character filtering module is the preprocessing step for videotext recognition, the performances of the other three modules (i.e., character separation, recognition, and refinement) are evaluated in the experiments. In the experiments, the kernel for the multi-class SVM classifier is the radial basis function (RBF) and the parameters are determined by threefold cross validation from the training data. In the character recognition module, k_G and k_R of (6) are assigned to 0.3 and 5, respectively.

First, character separation results are measured in terms of the character separation rate (CSR) which is defined as follows:

$$\mathrm{CSR} = \frac{T_S}{T} \tag{8}$$

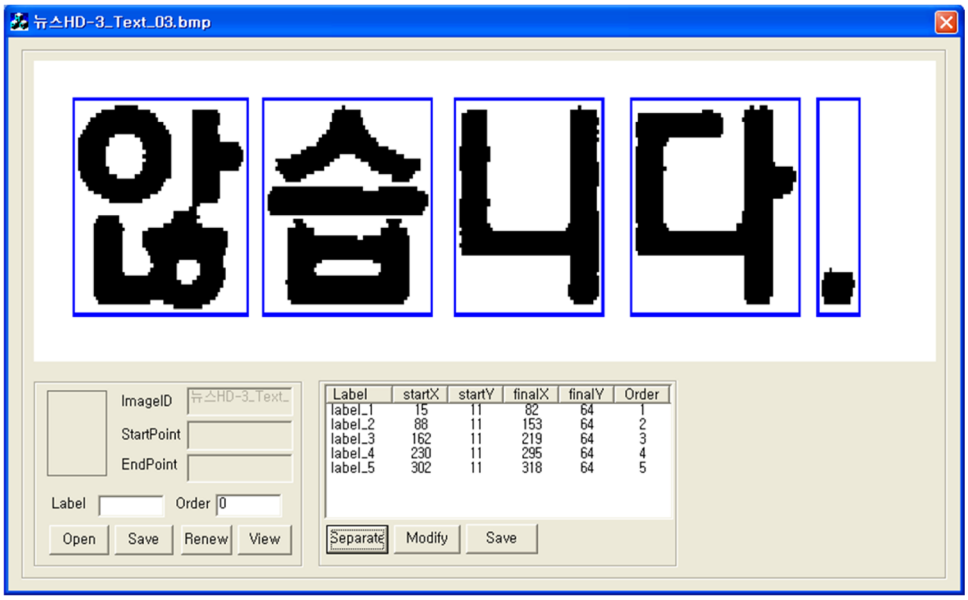

Fig. 11 The *Binarizer* to make the ground truth database for videotext recognition

Table 1 Example of the file structure in the ground truth database

Label	ImageID	StartX	StartY	EndX	EndY	Order
않	NewsHD-3_Text_03.bmp	15	11	82	64	1
습	NewsHD-3_Text_03.bmp	88	11	153	64	2
니	NewsHD-3_Text_03.bmp	162	11	219	64	3
다	NewsHD-3_Text_03.bmp	230	11	295	64	4
.	NewsHD-3_Text_03.bmp	302	11	318	64	5

Fig. 12 Some character separation and recognition results of text images. In each result, the first row shows a text image and its separation results while the second one shows the recognition result corresponding to the text image

where T_s denotes the number of characters separated correctly by the proposed method, and T denotes the total number of characters. Table 2 lists the CSR of the proposed method. As listed in the table, the CSR of the proposed method is 99.1 %.

Second, we evaluate the performance of the seven-type classification of characters in the character recognition module. Table 3 lists the performance evaluation results of the type classification. As can be seen, the total number of target class elements is 1,077, and the number of Korean characters is 1,000 ones among them. Here, the 1,000 ones are frequently used about 99 % of texts in the news videos. In Korean characters, the fourth type has the greatest number and is approximately 47 % of total class elements. The accuracy is evaluated as the relative frequency of the correctly classified types as follows:

$$\text{Accuracy} = \frac{\text{Number of correctly classfied types}}{\text{Number of characters}}. \quad (9)$$

The average rate of the type classification is 99.6 %. Notice that the accuracy of the sixth type is the lowest and this is because the characters of the sixth type have the most complex structure.

Third, the character recognition rate (CRR) of the character recognition module is measured. The CRR is defined as follows:

$$\text{CRR} = \frac{N_r}{N} \quad (10)$$

Table 2 Evaluation results of character separation

	Tr	T	CSR
Our method	175,221	176,884	0.991

Table 3 Evaluation results of type classification

Type	1st	2nd	3rd	4th	5th	6th	7th	Total
Number	118	71	46	469	260	36	77	1,077
Accuracy	0.998	0.993	0.981	0.997	0.999	0.970	0.999	0.996

where N_r denotes the number of characters recognized correctly by the proposed method, and N denotes the total number of characters. Figure 13 shows the evaluation results of the proposed character recognition method. As shown in the figure, the CRR of the proposed method is 95.1 %. Notice that the accuracy of the fourth type is the lowest and this seems to be due to the fact that the fourth type has the greatest number of character elements among the seven types. On the contrary, the sixth type is the highest and this is because the sixth type has the smallest number of class elements. Thus, the results show that the more the number of class elements is, the lower the recognition rate is.

Fourth, we evaluate the performance by character refinement. In the recognition results, most of recognition errors appear in the confusing pairs with small-size characters. It is because the shape ambiguity occurs in the case of small-size characters due to the character degradation and stroke distortion caused by video compression [19]. For example, there are "i" and "l", "I" and "T", "O" and "D", "면" and "연", "설" and "실", etc. Thus, the recognition results are refined by the character refinement module based on the language model, and some of the recognition errors in confusing pairs are corrected.

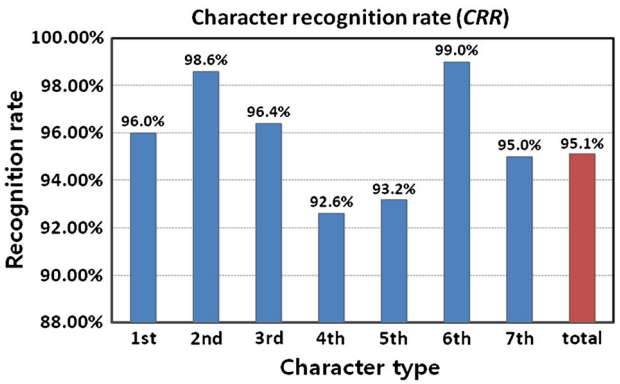

Fig. 13 Evaluation results of the character recognition (*CRR* character recognition rate). *Type 1–6* Korean characters. *Type 7* English characters

Figure 14 shows the recognition rate according to the top K candidates. The correct recognition result exists in the top K candidates. As shown in the figure, the recognition rate of the top ten candidates improves up to 98.9 %. That is, if the language model is applied to the character recognition results, the performance of the recognition rate can be improved. Table 4 shows the evaluation of the CRR for the intermediate results by each module in our method where two intermediate results are obtained: the result of the character separation and recognition, and the result after the character refinement which corresponds to the final result. As shown in the table, we improve the 1.1 % of the recognition rate by the character refinement module based on the language model.

3.3 Computational complexity

The proposed method has the merit of reducing computational complexity using geometric scores in the recognition. When we find the optimal segmentation path, the geometric scores including the squareness and internal gap are calculated. The path which has the low geometric score is eliminated in the recognition, and thus the computational overhead of recognition is reduced. To verify the efficiency in terms of the computational cost, we measure the time it takes to recognize videotexts without and with geometric scores, respectively. As shown in Table 5, the average processing time of the proposed method is 54 ms per character (ms/character) when geometric scores are employed for the recognition. Such results indicate that the

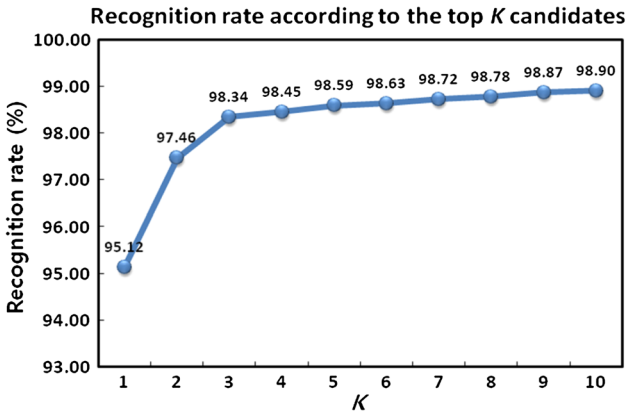

Fig. 14 Recognition rate according to the top K candidates (K: 1–10). The correct recognition result exists in the top K candidates

Table 4 CRR evaluation of intermediate results by each module

	Character recognition and separation	Character refinement
CRR	0.951	0.962

CRR character recognition rate

geometric scores are able to improve 25.0 % (18 ms/character) of the processing time without change of the recognition rate. In this case, the recognition rate is the same even after employing geometric scores.

3.4 Performance comparison

To see the recognition rates according to the resolution of video frames, we evaluate the performance on the text images from SDTV and HDTV videos, respectively. As shown in Table 6, the CRR of text images from HDTV videos is 98.3 and 2.3 % higher than that from SDTV videos. It is because text images in HDTV videos contain a comparatively small number of touching characters as compared with those in SDTV videos. In addition, the character degradation and stroke distortion caused by video compression are less serious in HDTV videos because of the high-resolution data. Furthermore, we compare the proposed method with [19] in terms of the CRR and computational time. Lee and Kim [19] has employed complementary combination of holistic and component analysis, and is one of the-state-of-the-art methods in the Korean videotext recognition. For a fair comparison, we test the two methods under the same circumstances, i.e., the same training and testing databases. As listed in Table 7, the proposed method produces nearly the same recognition rate as [19] in terms of CRR (the difference is only 0.3 %). Above all, the proposed method remarkably reduces the computational time by up to 94.8 % (994 ms/character) in comparison with [19]. Consequently, the experimental results demonstrate that the proposed method is very effective in the Korean-English bilingual videotext recognition in terms of the recognition rate as well as computational complexity.

Table 5 Computational time of the proposed videotext recognition method

	Without geometric scores	With geometric scores
Time	72	54

Experiments are performed on a PC with a Pentium IV 2.4 GHz and 2.00 GB memory using Visual C++ 6.0 based on the Window XP operating system. The unit of this test is ms per character (ms/character)

Table 6 CRR evaluation of text images by SDTV and HDTV videos

	SDTV	HDTV
CRR	0.960	0.983

CRR character recognition rate

Table 7 Comparison between the proposed method and Ref. [19] in terms of CRR and computational time

	CRR	Time
Proposed method	0.962	54
Ref. [19]	0.965	1,048

Experiments are performed on a PC with a Pentium IV 2.4 GHz and 2.00 GB memory using Visual C++ 6.0 based on the Window XP operating system. The unit of this test is ms per character (*ms*/character)

CRR character recognition rate

4 Conclusions

We have proposed a Korean-English bilingual videotext recognition method for news headline generation based on a split-merge strategy. Because special-characters have negative effects on the recognition performance, they are filtered before recognizing characters in the proposed method. Also, the proposed method employs a novel split-merge strategy for accurate character recognition. By the split-merge strategy, the text images are decomposed into several segments and the segments are merged into characters based on recognition scores. The recognition results are refined by the text refinement step based on the individual word language model. The final average recognition rate is 96.2 %, and thus our method is successfully applied to news headline generation. Experimental results and their analysis are reported in detail, thereby confirming that our method is capable of efficiently recognizing videotexts in news videos.

Although current character recognition method is sufficiently good, it still leaves quite a room for improvement of the recognition rate. In the future, we will continue improving the character recognition module using newly proposed classification technologies. In addition, we will further investigate the recommendation methods based on user preferences to provide personalized browsing and retrieval services to users.

Acknowledgments The partial work reported in this paper was conducted while the first author was with Samsung Electronics. The authors are grateful to Prof. Jinhyung Kim and Mr. Kyutae Cho in KAIST for their helpful discussion and the anonymous reviewers for their useful comments. This work was supported by the National Natural Science Foundation of China (Nos. 61050110144, 60803097, 60972148, 60971128, 60970066, 61072106, 61075041, 61003198, 61001206, and 61077009), the National Research Foundation for the Doctoral Program of Higher Education of China (No. 200807010003 and 20100203120005), the National Science and Technology Ministry of China (Nos. 9140A07011810DZ0107 and 9140A07021010DZ0131), the Key Project of Ministry of Education of China (No. 108115), and the Fundamental Research Funds for the Central Universities (Nos. JY10000902001, K50510020001, and JY10000902045).

References

1. Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., Jose, J.M.: Video browsing interfaces and applications: a review. SPIE Rev. **1**, 018004 (2010)
2. Lee, C.C., Shih, C.Y., Huang, H.M.: Story-related caption detection and localization in news video. Opt. Eng. **48**, 037005 (2009)
3. Dimitrova, N., Zhang, H.J., Shahraray, B., Sezan, I., Zakhor, A., Huang, T.: Applications of video content analysis and retrieval. IEEE Multimed. **9**, 43–55 (2002)
4. Dimitrova, N., McGee, T., Elenbaas, H.: Video key-frame extraction and filtering: a key-frame is not a key-frame to everyone. In: Proceedings of ACM International Conference on Knowledge and Information Management, pp. 113–120 (1997)
5. Jasinschi, R. S., Dimitrova, N., McGee, T., Agnihotri, L., Zimmerman, J., Li, D.: Integrated multimedia processing for topic segmentation and classification. In: Proceedings of IEEE International Conference on Image Processing, pp. 366–369 (2001)
6. Kim, J.G., Chang, H.S., Kang, K., Kim, M., Kim, J., Kim, H.M.: Summarization of news video and its description for content-based access. Int. J. Imaging Syst. Technol. **13**, 267–274 (2003)
7. Merialdo, B., Lee, K.T., Luparello, D., Roudaire, J.: Automatic construction of personalized TV news program. In: Proceedings of ACM International Conference on Multimedia, pp. 323–331 (1999)
8. Liu, J., He, Y., Peng, M.: NewsBR: a content-based news video browsing and retrieval system. In: Proceedings of Computer and Information Technology, pp. 857–863 (2004)
9. Kim, S.K., Hwang, D.S., Kim, J.Y., Seo, Y.S.: An effective news anchorperson shot detection method based on adaptive audio/visual method generation. Lect. Notes Comput. Sci. **3568**, 276–285 (2005)
10. Gao, X., Li, J., Yang, B.: A graph-theoretical clustering based anchor person shot detection for news video indexing. In: Proceedings of International Conference on Computational Intelligence and Multimedia Applications, pp. 108–113 (2003)
11. Zhu, W., Toklu, C., Liou, S.P.: Automatic news video segmentation and categorization based on closed-captioned text. In: Proceedings of IEEE International Conference on Multimedia and Expo, pp. 1036–1039 (2001)
12. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition, pp. 511–518 (2001)
13. Jung, C., Liu, Q., Kim, J.K.: A new approach for text segmentation using a stroke filter. Signal Process. **88**, 1907–1916 (2008)
14. Jung, C., Liu, Q., Kim, J.K.: Accurate text localization in images based on SVM output scores. Image Vis. Comput. **27**, 1295–1301 (2009)
15. Jung, C., Liu, Q., Kim, J.K.: A stroke filter and its application to text localization. Pattern Recogn. Lett. **30**, 114–122 (2009)
16. Sato, T., Kanade, T., Highes, E.K., Smith, M.A.: Video OCR for digital news archive. In: Proceedings of IEEE Workshop on Content-Based Access of Image and Video Database, pp. 52–60 (1998)
17. Sato, T., Kanade, T., Hughes, E.K., Smith, M.A., Satoh, S.: Video OCR: indexing digital news libraries by recognition of superimposed captions. Multimed. Syst. **7**, 385–395 (1999)
18. Chang, F., Chen, G.C., Lin, C.C., Lin, W.H.: Caption analysis and recognition for building video indexing systems. Multimed. Syst. **10**, 344–355 (2005)
19. Lee, S., Kim, J.: Complementary combination of holistic and component analysis for recognition of low-resolution video character image. Pattern Recogn. Lett. **29**, 383–391 (2008)
20. Wang, F., Ngo, C.W., Pong, T.C.: Structuring low-quality videotaped lectures for cross-reference browsing by video text analysis. Pattern Recogn. **41**, 3257–3269 (2008)
21. Park, J., Lee, G., Kim, E., Lim, J., Kim, S., Yang, H., Lee, M., Hwang, S.: Automatic detection and recognition of Korean text in outdoor signboard images. Pattern Recogn. Lett. **31**, 1728–1739 (2010)
22. Chang, Y., Chen, D., Zhang, Y., Yang, J.: An image-based automatic Arabic translation system. Pattern Recogn. **42**, 2127–2134 (2009)
23. Wolf, C., Jolion, J.M.: Extraction and recognition of artificial text in multimedia documents. Pattern Anal. Appl. **6**, 309–326 (2003)
24. Chen, D., Odobez, J.M., Bourlard, H.: Text detection and recognition in images and video frames. Pattern Recogn. **13**, 595–608 (2004)
25. Tang, X., Gao, X., Liu, J., Zhang, H.: A spatio-temporal approach for video caption detection and recognition. IEEE Trans. Neural Netw. **13**, 961–971 (2002)
26. Lienhart, R., Wernicke, A.: Localizing and segmenting text in images and videos. IEEE Trans. Circuit Syst. Video Technol. **12**, 256–267 (2002)
27. Yang, H., Siebert, M., Lühne, P., Sack, H., Meinel, C.: Automatic lecture video indexing using video OCR technology. In: Proceedings of IEEE International Symposium on Multimedia, pp. 111–116 (2011)
28. Sarfraz, M.S., Shahzad, A., Elahi, M.A., Fraz, M.: Real-time automatic license plate recognition for CCTV forensic applications. J. Real Time Image Process. (2011). doi:10.1007/s11554-011-0232-7
29. Chin, S., Choi, Y., Choo, M.: A skew free Korean character recognition system for PDA devices. In: Proceedings of International Conference on Intelligent Computing, pp. 476–483 (2006)
30. Sharma, N., Pal, U., Blumenstein, M.: Recent advances in video based document processing: a review. In: Proceedings of IAPR International Workshop on Document Analysis Systems, pp. 63–68 (2012)
31. Kim, M.S., Cho, K.T., Kwag, H.K., Kim, J.H.: Segmentation of handwritten characters for digitalizing Korean historical documents. In: Proceedings of International Conference on Document Analysis and Recognition, pp. 114–124 (2004)
32. Tseng, Y.H., Lee, H.J.: Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm. Pattern Recogn. Lett. **20**, 791–806 (1999)
33. Kang, K.W., Kim, J.H.: Utilization of hierarchical, stochastic relationship modeling for Hangul character recognition. IEEE Trans. Pattern Recogn. Mach. Intell. **26**, 1185–1195 (2004)
34. Kim, J.H., Kim, K.K., Chien, S.I.: Korean and English character recognition system using hierarchical classification neural network. In: Proceedings of IEEE International Conference on Systems, Man and Cybernetics, pp. 759–764 (1995)
35. Lim, K.T.: A study on machine printed character recognition based on character type classification. J. Electron. Eng. Korea **40**, 26–39 (2003)
36. Kwak, N., Choi, C.H.: Input feature selection by mutual information based on Parzen window. IEEE Trans. Pattern Recogn. Mach. Intell. **24**, 1667–1771 (2002)
37. Fisher, R.A.: The statistical utilization of multiple measurements. Ann. Eugen. **8**, 376–386 (1938)
38. Ryu, S., Kim, J.H.: A language model using variable length tokens for open-vocabulary Hangul text recognition. Pattern Recogn. **37**, 1549–1552 (2004)
39. Ryu, S., Kim, J.H.: Learning the lexicon from raw texts for open-vocabulary Korean word recognition. In: Proceedings of International Conference on Document Analysis and Recognition, pp. 202–206 (2003)

40. Bagdanov, A., Kanai, J.: Projection profile based skew estimation algorithm for JBIG compressed images. In: Proceedings of International Conference on Document Analysis and Recognition, pp. 401–405 (1997)

Author Biographies

Cheolkon Jung received the BS, MS, and Ph.D. degrees in Electronic Engineering from Sungkyunkwan University, Republic of Korea, in 1995, 1997, and 2002, respectively. He was with the Samsung Advanced Institute of Technology (Samsung Electronics), Republic of Korea, as a research staff member from 2002 to 2007. He was a research professor in the School of Information and Communication Engineering at Sungkyunkwan University, Republic of Korea, from 2007 to 2009. Since 2009, he has been with the School of Electronic Engineering at Xidian University, China, where he is currently a professor. His main research interests include computer vision, pattern recognition, image and video processing, multimedia content analysis and management, and 3DTV.

Licheng Jiao received the BS degree from Shanghai Jiao Tong University, China, in 1982, and the MS and Ph.D. degrees from Xi'an Jiao Tong University, China, in 1984 and 1990, respectively. From 1990 to 1991, he was a Postdoctoral Fellow in the National Key Lab for Radar Signal Processing at Xidian University, China. Since 1992, he has been with the School of Electronic Engineering at Xidian University, China, where he is currently a distinguished professor. He is also the Dean of the School of Electronic Engineering and the Institute of Intelligent Information Processing at Xidian University, China. His current research interests include signal and image processing, nonlinear circuit and systems theory, learning theory and algorithms, computational vision, computational neuroscience, optimization problems, wavelet theory, data mining, and 3DTV.