2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Recognition and Transition Frame Detection of Arabic News Captions for Video Retrieval

Seiya Iwata, Wataru Ohyama, Tetsushi Wakabayashi and Fumitaka Kimura
Graduate School of Engineering, Mie University
1577 Kurimamachiya-cho, Tsu-shi, Mie 514-8507, Japan
Email: [iwata, ohyama, waka]@hi.info.mie-u.ac.jp

***Abstract*—The authors have conducted studies on recognizing Arabic news captions to develop a system for video retrieval to index and edit Arabic broadcast programs daily received and stored in big database. This paper describes a dedicated OCR for recognizing low resolution news captions in video images. News caption recognition system consisting of text line extraction, word segmentation and segmentation-recognition of words is developed and the performance was experimentally evaluated using datasets of frame images extracted from AlJazeera broadcasting programs. Character recognition of moving news caption is difficult due to combing noise yielded by the interlacing of scan lines. A technique to detect and eliminate the combing noise to correctly recognize the moving news caption is proposed. This paper also proposes a technique based on inter-frame text difference to detect transition frame of still news captions. The technique to detect transition frames is necessary for efficient video retrieve and play. The proposed technique is experimentally tested and shown to be robust to quick motion of the background and is able to detect the transition frame correctly with the F-measure higher than 90%. When compared with the ABBY FineReader 11®commercial OCR the dedicated OCR improves the recall of the Arabic characters in AlJazeera broadcasting news from 70.74% to 95.85% for non-interlaced moving news captions and from 23.82% to 96.29% for interlaced moving news captions.**

***keywords*—OCR, News caption recognition, Arabic word recognition, Combing noise, Video retrieval, Moving news caption.**

I. Introduction

Analog satellite broadcasting is widely used for international Arabic TV broadcasting and is an important news source of Arabic and Middle East area. The authors have studied on Arabic news caption recognition to develop a system for video retrieval by keyword to extract and edit Arabic broadcast programs daily received and stored in a big database. There are two approaches for keyword retrieval of document image. One is to use general keyword retrieval for OCR output of the document image, and the other is word spotting based approach that detects text regions having similar shape to the input keyword. The former enables us high speed full text retrieval with wide range of applications and is suitable for relatively good quality OCR readable document. The latter is suitable for such documents as low resolution, historical and/or handwritten documents.

This paper describes a dedicated OCR for recognizing low resolution news caption in video images to generate the text for full text retrieval and other applications such as automatic language translation. This paper also proposes a technique based on inter-frame text difference to detect transition frame of still news captions.

The system requires two basic functions of text extraction from the video image and the character recognition of the extracted text image.

Many papers have been reported in the area of text extraction in video documents [1].

Lyu et al. proposed a region based sequential multi-resolution paradigm, in which no text edges can appear several times at different resolution levels [2]. The edge map is created by Sobel filtering and a local thresholding operation and horizontal/vertical projections are used to locate the text objects. English and Chinese news are tested and the recall and precision of the text detection were 91.1% and 90.8% respectively.

Dubey proposed a region base combined technique of vertical edge detection by Sobel filtering and Accumulative Intensity Morphing (AIM) [3]. Following the edge detection region grouping is applied by using the AIM, which connects start pixels and end pixels horizontally. Then a number of horizontal-adjacent vertical lines satisfying specified conditions are considered as candidate text regions. The recall and precision for captured image from TV programs were 91.6% and 86% respectively.

Gllavata et al. presented a temporal-redundancy-based method for detecting text object in videos. A Fuzzy Clustering Ensemble (FCE) is adopted to fuse multi-frame information [4]. The features are extracted by wavelet transform and clustered by Fuzzy C Means (FCM). English videos are tested and the recall and precision of the text detection were 92.04% and 96.71% respectively.

Lefevre et al. presented a combined technique of the region base and texture base approach [5]. Color-related detector, wavelet-based texture detector [6], edge-based detector [7] [8], and temporal invariant principle are used to detect candidate caption regions. News and commercial videos including 322 texts were tested and the recall and precision were 92% (60%) and 76% (90%) respectively.

The above survey shows that the recall of the text extraction for video image is around 92%.

Meanwhile the recall and precision of the character recognition of the extracted Arabic text region were 91.78% and 91.72% respectively when evaluated using ABBY FineReader 11®[9] commercial OCR (detailed in section III-F). The OCR

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

(a) Input image

(b) Binary image

(c) Text line detection

(d) Word detection

```
؟؟؟؟؟ قوات من الجيش المصري تشرع في هدم منازل في
رفح تمهيدا لإقامة منطقة عازلة مع قطاع غزة
مكة | 07:01 إلجيش المصري يبدأ هدم منازل مواطنين برفح ضمن المنطقة العازلة المزمع ا
```

(e) Result of word recognition

Fig. 1: Outline of detection and recognition of moving news caption

performance is considerably lower than the one for general scanned document images because of the lack of resolution and quality of the video images. If the character level recall is 92% the recall of the word with five characters will be 66%, hence the recall of the keyword retrieval will also be 66%.

While many researches have been done for detection and recognition of text for video image retrieval they are still on the way of the performance improvement. The purpose of this paper is to develop a dedicated OCR for recognizing low resolution news caption in video images to improve the performance of the character recognition and the keyword retrieval.

Character recognition of moving news caption is deteriorated by combing noise yielded by the interlacing of scan lines.

The character level recall of the above commercial OCR for such moving news caption was 44.13%. This paper proposes a technique to detect and eliminate the combing noise to correctly recognize the moving news caption. The study on transition detection of still news captions based on inter-frame text difference has not been studied so far to the best of our knowledge.

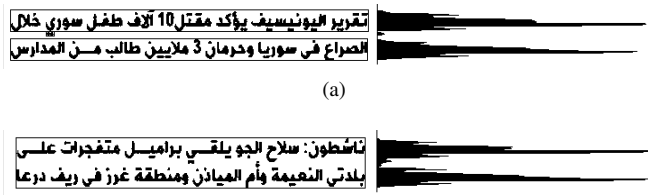
II. News caption recognition procedure

A. Outline of news caption recognition

The outline of news caption recognition is as follows.

1) Binarization of input image (Fig. 1(a), (b)).
2) Elimination of connected components with width and height greater than thresholds.
3) Text line detection by vertical profile analysis (Fig. 1(c)).
4) Combing noise elimination of moving news caption by shifting odd number of scan lines to left by a pixel.
5) Word segmentation of text lines by thresholding gaps between connected components (Fig. 1(d)).
6) Repeat steps 2) to 5) to the image with reversed foreground and background.
7) Segmentation and recognition of characters in a word (Fig. 1(e)).
8) Transition frame detection of still news captions in a sequence of frames.

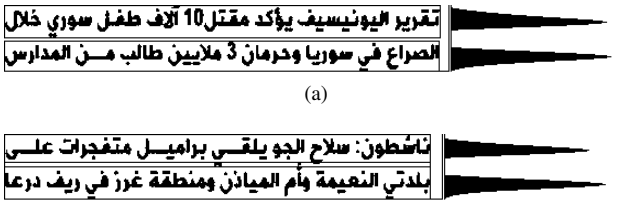
Details of each step is described in the following subsections.



(a)

(b)

Fig. 2: Over separation of diacritics (a) and adherence of text lines (b)



(a)

(b)

Fig. 3: Text lines detected by 1-dimensional DOG filtered vertical projection and fixed neighbor classification rule

B. Text line detection

1) Binarization of input image (Fig. 1(b)): The threshold is determined by Otsu's method [10] and the input image (frame) is binarized by the threshold.

2) Elimination of unnecessary connected components: Let the width and height of the input image be denoted by W and H, and those of a connected component by w and h. Then the connected components with $w > W/4$ or $h > H/8$ are eliminated as non-text line components.

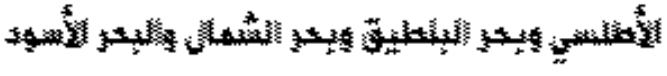
3) Text line detection by vertical profile analysis (Fig. 1(c)): A frame image often contains multiple-line still news captions with a moving news caption.

The vertical profile analysis based news caption line detection encounters such problems as over separation of diacritics above or below Arabic letters (Fig. 2(a)), and adherence of lines due to poor separability of the vertical profile (Fig. 2(b)). In order to solve these problems 1-dimensional difference of Gaussian filter (DOG) is applied to the vertical profile. Black pixels in positive region of the DOG filtered vertical projection are extracted as text lines and black pixels in negative region are classified to its nearest text lines by fixed neighbor hood classification rule (Fig. 1(c), Fig. 3(a), (b)). The vertical profiles in Fig. 1(c) and Fig. 3(a), (b) show positive part of the DOG filtered projections. The variances of the two Gaussians in DOG filter are 5 and 200.

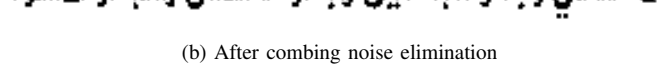
4) Elimination of combing noise of moving news caption: If the total number of left edge cavity and right edge spur shown in Fig. 4 is greater than a threshold the text line is classified as moving news caption otherwise classified as still news caption. Combing noise of the moving news caption is eliminated by shifting odd scan lines to left one pixel (Fig. 5(b)).

0	1	1	1	0	0
0	0	1	1	1	0
0	1	1	1	0	0

Fig. 4: Left edge cavity and right edge spur



(a) Before combing noise elimination



(b) After combing noise elimination

Fig. 5: Elimination of combing noise

5) False text line reduction: The more the false text lines are erroneously detected from the background image, the more the successive processing of word segmentation and recognition is required. False text line reduction is performed aiming to increase the processing speed without sacrificing the recall of words and characters. To easily detect and remove false lines the average of the eccentricity of a connected component

$$e = \frac{perimeter^2}{area}, \tag{1}$$

is calculated for all components in the text line and the line is removed if the average is less than 30.

Then the vertical position (interval) of the baseline is approximated by the interval where the vertical projection is greater than 80% of the maximum projection. The text line is removed if the width of the baseline is greater than 30% of the height of the line. Furthermore crossing counts (number of changes from white pixel to black pixel) above, in and below the baseline, N_A, N_I, N_B are calculated and the text line is removed if the next conditions hold.

$$N_A < N_I \quad or \quad N_A < N_B \tag{2}$$

By the above processing are removed 8 false text lines shown in Fig1(c), and remaining 3 lines are fed to the next word segmentation task (Fig1(d)).

C. Word segmentation

Arabic words are generally separated by spaces. The space detection is carried out by thresholding operation to classify horizontal gaps between connected components to between word gap and within word gap (Fig. 1(d), Fig. 6). The threshold is determined by Otsu's discriminant analysis based method [10] for each text line.

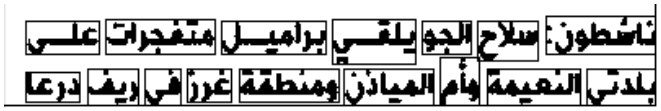


Fig. 6: Result of word segmentation

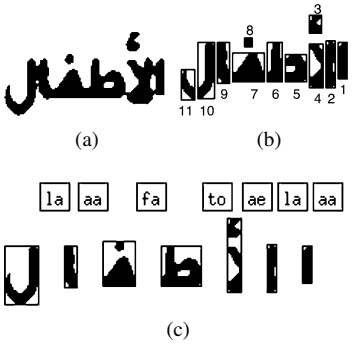


Fig. 7: Input word image (a), primitive segments (b) and character segmentation (c)

D. Recognition of characters and words

Character recognition is carried out using 64-dimensional feature vector of chain code histogram and the modified quadratic discriminant function (MQDF)[11][12].

1) Over segmentation of characters: Since it is difficult to correctly segment characters without recognizing them, the characters are first over segmented and then merged into each character in the process of character recognition.

The segmentation points are detected through local extrema analysis of the upper contour of the word image. Among the local minima (valley points), those that are not deep enough from the adjacent local maxima are sequentially removed. Then the black pixels in the column including the segmentation point are replaced by white pixels and the connected components are detected as primitive segments with their enclosing rectangles (Fig. 7(b)).

2) Word recognition by dynamic programming: The number of the primitive segments is usually greater than that of the characters in the word. In order to merge these primitive segments into characters so that the final character segmentation is optimal, dynamic programming (DP) is applied using the total likelihood of characters as the objective function.

To apply the DP technique, the primitive segments are sorted right to left according to the centroids of the enclosing rectangles. If two or more rectangles have the same x-coordinates of the centroids they are sorted top to bottom. Numbers in Fig. 7(b) shows the order the sorted primitive segments. Rectangles 1 in Fig. 7(b) corresponds to the letter 'aa' of aa/la/ae/to/fa/aa/la, rectangle 2 to 'la' and rectangles 3 - 4 to 'ae' ... and so on.

These assignments of rectangles to letters can be represented as TABLE I, where i denotes the letter number, $j(i)$ denotes the number of the last rectangle corresponding to the i-th letter.

TABLE I: Assignment of letters

i	1	2	3	4	5	6	7
${A_i}^*$	aa	la	ae	to	fa	aa	la
$j(i)$	1	2	4	6	8	9	11

Note that the number of the first rectangle corresponding to the i-th letter is $j(i-1)+1$.

Given $j(i)$, $i = 1, 2, ..., n$, the total likelihood of the character is defined by

$$\begin{aligned} L &= \sum_{i=1}^{n} l\big({A_i}^*, j(i-1)+1, j(i)\big), \\ &= \sum_{i=1}^{n} \max_{A_i} \big\{ l\big(A_i, j(i-1)+1, j(i)\big) \big\}, \end{aligned} \tag{3}$$

and the output word is represented by ${A_1}^*{A_2}^*...{A_n}^*$, where n is the number of letters and $l(A_i, j_1, j_2)$ is the likelihood that the segments j_1 to j_2 belongs to letter A_i. The optimal assignment $j(i)^*$ $(i = 1, 2, ..., n)$ which maximizes the total likelihood is found using the dynamic programming. Since the number of letters n is unknown it is estimated to be $m/2 \leq n \leq m$, where m is the number of the primitive segments. After the optimal assignments for different n in the estimated interval are found, the word with maximum average likelihood per letter L^*/n is selected as the output of the word recognition.

3) False word reduction: To reduce the false word is performed the thresholding operation for the maximum average likelihood L^*/n. The threshold works as a parameter to trade off the recall and precision of the words. One false word detected in Fig.1(d) is removed by the above processing and are shown by "??????" in the output (Fig.1(e)).

4) Transition frame detection of still news captions: Proposed technique uses inter-frame text difference to detect transition frame of still news captions. This technique is robust to quick motion of the background and is able to detect the transition frame correctly. The method proposed in this paper is as follows.

1) Detect the operations of insertion, deletion and substitution with minimum edit distance to convert the preceding news caption into the succeeding news caption. Where the edit distance is the total number of the insertions, deletions and substitutions and is minimized by dynamic programming technique (Vierbi algorithm).
2) f-measure of the inter-frame text difference is defined as follows,

$$\begin{aligned} r &= (M_1 - N_D - N_S)/M_1 \\ p &= (M_2 - N_I - N_S)/M_2 \\ f &= 2rp/(r+p), \end{aligned} \tag{4}$$

where N_I, N_D, N_S are the number of insertion, deletion and substitution and M_1, M_2 are number of characters in preceding news caption and succeeding news caption respectively.

TABLE II: Correct recognition rate of characters in news captions in dataset I (%)

	Five-fold cross validation	Re-substitution method
Still news caption	99.18(5332/5376)	99.65(5357/5376)
Moving news caption	98.69(6465/6551)	99.42(6513/6551)

3) If the above f-measure falls below a threshold the succeeding frame is determined to be the transition frame. OCR output of the news captions usually involves segmentation error or recognition error of characters. Successive two frames involving moving news captions yield a certain amount of inter-frame difference even though none of them is a transition frame.
The above threshold has to be determined not to detect false transition frame due to the small inter-text difference.

III. Experiment

A. Datasets

Proposed news caption recognition system is experimentally tested using 161 frame images extracted from AlJazeera satellite broadcasting TV programs. This dataset (dataset I) is divided into five subsets to perform five-fold cross validation test. When a subset is tested the rest of four subsets are used to train the character classifier. The frame images in dataset I were extracted with arbitrary (variable) time period over about 1 hour news programs. As another independent dataset (dataset II), 292 frame images were extracted from AlJazeera Livestation program downloaded via the internet. This dataset is used for comparative performance evaluation with other technique. The frame images in dataset II were extracted with constant time period of 1 second over about 5minutes news program.

The frame images in dataset I are interlaced and those in dataset II are non-interlaced.

B. Character recognition test for dataset I

Character recognition test is performed for 161 frame images in dataset I. TABLE II shows correct recognition rate of characters in news captions extracted from the dataset I. They are 99.18% for still news caption and 98.69% for moving news caption when evaluated by five-fold cross validation. The total number of characters is 5376 and 6551 in each case respectively.

C. Word recognition test for dataset I

Word recognition test is performed for 161 frame images in dataset I. TABLE III shows correct recognition rate of words in news captions extracted from the dataset I. The word recognition rates are 94.29% for still news caption and 95.28% for moving news caption when evaluated by five-fold cross validation. The total number of words is 1138 and 1336 in each case respectively.

TABLE III: Correct recognition rate of word in news captions in dataset I (%)

	Five-fold cross validation	Re-substitution method
Still news caption	94.29(1073/1138)	95.08(1082/1138)
Moving news caption	95.28(1273/1336)	96.93(1295/1336)

TABLE IV: Performance of the transition frame detection of still news captions in dataset I (%)

Threshold	$Recall$	$Precision$	$F-measure$
50	66.43(93/140)	98.94(93/94)	79.49
60	87.86(123/140)	99.19(123/124)	93.18
70	89.29(125/140)	96.15(125/130)	92.59
80	90.71(127/140)	95.49(127/133)	93.04
90	90.71(127/140)	94.78(127/134)	92.70

TABLE V: Performance of the transition frame detection of still news captions in dataset II (%)

Threshold	$Recall$	$Precision$	$F-measure$
50	0 (0/6)	100.0(0/0)	2.00
60	50.00(3/6)	100.0(3/3)	66.67
70	83.33(5/6)	100.0(5/5)	90.91
80	100.0(6/6)	54.55(6/11)	70.59
90	100.0(6/6)	6.52(6/92)	12.24

D. Transition frame detection of still news captions

The performance of the transition frame detection of still news captions in dataset I is experimentally evaluated.

Here, the recall R, precision P and F-measure F are defined as follows:

$$\begin{aligned} R &= A/B \\ P &= A/C \\ F &= 2RP/(R+P), \end{aligned} \tag{5}$$

where A, B and C denote the number of frames both in true transition frames and detected frames, in true transition frames and in detected frames respectively. The frame with no still news caption is assumed to have empty news caption. There are 140 transition frames out of 161 frames.

TABLE IV shows recall, precision and F-measure of the transition frame detection. In the next evaluation test, the OCR outputs of the news captions in dataset II are used for the transition frame detection. There are 6 transition frame out of 292 frames. TABLE V shows recall, precision and F-measure of the transition frame detection.

E. Processing time

The average processing time per frame in dataset I is 125.8 msec shown in TABLE VI in detail. Used CPU is Intel(R) Core(TM)2 Quad CPU 2.66Gz and video resolution is 720×480.

F. Comparison with other method

TABLE VII shows the Arabic text recognition performance of ABBY FineReader 11®[9] commercial OCR. The test data

TABLE VI: Average processing time per frame in dataset I (msec)

Line detection	Word detection	Word recognition	Transition frame detection	Total
30.4	17.1	78.1	0.2	125.8

TABLE VII: Performance of commercial OCR for moving news captions in dataset I and dataset II(%)

Used data	Text extraction	*Recall*	*Precision*	$F-measure$
Data set I (Interlacing)	Manual	44.13	53.44	48.34
	Auto	23.82	47.27	31.68
Data set II (Non-interlacing)	Manual	91.78	91.72	91.75
	Auto	70.74	72.02	70.99

TABLE VIII: Performance of proposed OCR for moving news captions in dataset I and dataset II)(%)

Used data	Text extraction	*Recall*	*Precision*	$F-measure$
Data set I (Interlacing)	Auto	96.29	95.62	95.95
Data set II (Non-interlacing)	Auto	95.86	95.04	95.45

is the text region of the moving news captions in dataset I and dataset II. The recall and precision for manually extracted text regions in the dataset I with interlacing were 44.13% and 53.4% respectively and were 23.82% and 47.27% for entire frame image (without manual text extraction). The performance of the OCR is severely deteriorated by the combing noise due to the interlacing scan. The recall and precision for manually extracted text regions in the dataset II without interlacing were 91.78% and 91.72% respectively and were 70.74% and 72.02% for entire frame image. The recall is decreased by about 20% because of the failure of the automatic text extraction. Even the OCR performance for manually extracted text regions is considerably lower than the one for general scanned document images because of the lack of resolution and quality of the video images. TABLE VIII shows the performance of the dedicated OCR described in this paper. The performance is significantly improved for both of the dataset.

IV. Conclusion

News caption recognition system consisting of text line extraction, word extraction and segmentation-recognition of words was developed and the performance was experimentally evaluated using datasets of frame images extracted from AlJazeera broadcasting programs. This paper proposes a technique to detect and eliminate the combing noise to correctly recognize the moving news caption.

When compared with ABBY FineReader 11®commercial OCR the dedicated OCR improves the recall of the Arabic character recognition from 70.74% to 95.85% for non-interlaced moving news caption images and from 23.82% to 96.29% for interlaced moving news caption images.

This paper also proposed a technique based on inter-frame text difference to detect transition frame of still news captions. The technique was experimentally tested and shown to be robust to quick motion of the background and was able to detect the transition frame correctly with the F-measure higher than 90%.

Followings are remaining as future research topics.

1) Accuracy improvement of word recognition and the transition frame detection.
2) Connecting moving news captions to reconstruct the sentence.
3) Application to other Arabic and Persian broadcasting programs.

References

[1] J. Zhang and R. Kasturi, "Extraction of Text Objects in Video Documents: Recent Progress", Proceedings of IAPR International Workshop on Document Analysis Systems (DAS), pp. 5-17, 2008.

[2] M. R. Lyu, J. Song, and M. Cai, "A Comprehensive method for multilingual video text detection, localization, and extraction", IEEE transactions on circuits and systems for video technology, Vol. 15, pp. 243-255, 2005.

[3] P. Dubey, "Edge Based Text Detection for Multi-purpose Application", Proceedings of International Conference on Signal Processing, Vol. 4, 2006.

[4] J. Gllavata, E. Qeli, and B. Freisleben, "Detecting Text in Videos Using Fuzzy Clustering Ensembles", Proceedings of the Eighth IEEE International Symposium on Multimedia, pp.283-290, 2006.

[5] S. Lefevre and N. Vincent, "Caption localization in video sequences by fusion of multiple detectors", Proceedings of eighth International Conference on Document Analysis and Recognition (ICDAR), pp. 106-110, 2005.

[6] H. Li, D. Doerman, and O. Kia, "Automatic text detection and tracking in digital video", IEEE Transactions on Image Processing, 9(1):147-156, 2000.

[7] C.Wolf and J. Jolion, "Extraction and recognition of artificial text in multimedia documents", Pattern Analysis and Applications, 6:309-326, 2003.

[8] S. Lefevre, C. Dixon, C. Jeusse, and N. Vincent, "A local approach for fast line detection", In IEEE International Conference on Digital Signal Processing, volume 2, pages 1109-1112, Santorini, Greece, August 2002.

[9] "OCR software—FineReader 11", http://FineReader.add-soft.jp/, (accessed 2014-5-25)

[10] N. Otsu, "A threshold selection method from gray-level histogram", IEEE Trans. Systems, Man and Cybernetics, vol. SMC-9, pp. 62-69, 1979.

[11] F. Kimura, K. Takashina, S. Tsuruoka, and Y. Miyake, "Modified quadratic discriminant functions and the application to Chinese character recognition", IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-9, no. 1, pp. 149153, Jan 1987.

[12] F. Kimura, M. Shridhar, and Z. Chen, "Improvements of a Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words", Proc. of the Second International Conference on Document Analysis and Recognition (ICDAR), Tsukuba, pp.18-22, 1993.