Overlay Text Detection and Recognition for Soccer Game Indexing

J. Ngernplubpla and O. Chitsophuk, *Member, IACSIT*

***Abstract*—In this paper, new multiresolution overlaid text detection and recognition is proposed to detect and recognize the text characters in various illuminations and complex backgrounds. The proposed text detection algorithm is based on heuristic top-down, which detects candidate text regions at the top level of the pyramid using the proposed adaptive threshold and noise reduction. The multiresolution detection allows us to reduce noise and unwanted regions in the video frame while the proposed adaptive threshold helps to automatically select a suitable threshold for various illumination conditions. Once the bounding boxes of the text regions have been extracted, they will be enhanced and segmented into character regions. The segmented characters are recognized using optical character recognition. For the low quality of the video frame, it may results in poor character recognition rate. The recognition precision is about 63.88%. Since the performance of OCR systems closely relies on the quality of the targeting scene, in this paper, we proposed the keyword matching based on character refinement (KMCR), which uses interpolation search to find the matches. These matched keywords are temporally tracked and voted for the final results. The KMCR helps to increase the accuracy and reliability to the recognition results, thus leads to an improvement on the video indexing.**

***Index Terms*—Text detection, text recognition, text enhancement, adaptive threshold for sobel edge detection, background subtraction, multiresolution.**

I. Introduction

With the advance in multimedia, communications and storage technologies, video becomes one of the most popular types of media in various applications. In order to effectively utilize the video repositories, the semantic video analysis and management for video understanding, indexing, and retrieval are necessary. Text embedded in images and video sequences generally provides brief and important content information, such as the name of a player or speaker, the title, location, date, or important events, etc. For example, in a soccer video, the overlaid text can be used to identify the interesting events such as goal, free kick, yellow card, and red card, which will be benefit for producing special video summaries. There are two types of appearing text in a video, Scene and Graphic/Overlaid text [4]. The scene text naturally occurs in the recording scene while the overlaid text is superimposed onto the video using title generators [5]. Scene text is usually written on some objects. The size is big but the shape can be deformed, occluded, or positioned in various alignments and movements. However, the overlaid text is not occluded but small size and located in complex background. In this paper, we focus on the overlaid text detection and recognition. Even though the overlaid text in complex background is difficult to detect and recognize, it is typically very relevant to the associated images or video frames.

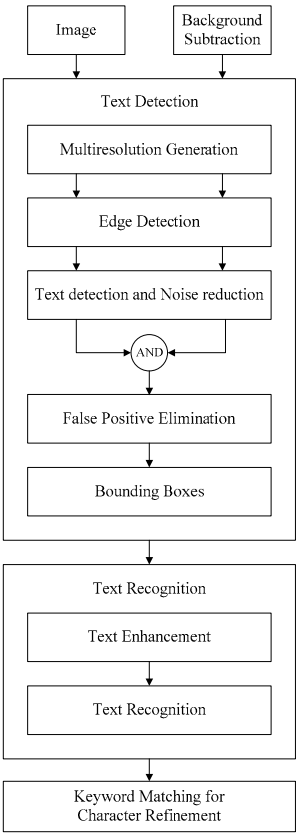


Fig. 1. Overview of the proposed system

Several researches have been proposed for text detection in the video frame [1], [2], [6]. In [6], the authors proposed text detection using Laplacian window for edge detection and using Maximum Gradient Difference for connected component. After that, K-mean clustering is used to classify the detected text regions into actual text or background regions. The text detection techniques based on edge detection and morphological noise reduction are proposed in [1], [2]. In [2], text detection was proposed based on multiresolution processing and canny edge detection, which allowed the algorithm to be able to detect various sizes of text. The text detection algorithm in [1] was based on sobel edge detection. The candidate text regions were recognized using the commercial OCR then keyword matching was performed to detect a keyword (or keywords) that match with soccer concept(s) from their term-database. Even though those algorithms seem to be efficient for text detection, it was not able to detect the overlaid text in the complex background

Manuscript received April 12, 2012; revised May 27, 2012.

The authors are with King Mongkut's Institute of technology Ladkrabang, Bangkok, Thailand (e-mail: thnredline@ gmail.com, kcoracha@kmatl.ac.th).

with various illumination changes in a single scene or throughout a sequence of frames and scene with large amount of noises.

Therefore, in this paper, new multiresolution overlaid text detection and recognition is proposed. The proposed text detection algorithm is based on heuristic top-down, which detects candidate text regions at the top level of the pyramid using the proposed adaptive threshold and noise reduction then the detected text regions will be segmented and mapped backward onto the original resolution. Each segmented text line is further partitioned into character sub-regions, which will be recognized using commercial OCR software. Due to the very low resolution of the characters, the OCR software is not able to achieve good recognition results. Therefore, the keyword matching based on character refinement (KMCR) is studied for improving the final recognition text string by searching for meaningful keywords from the database and voting the matched keywords through temporal frames.

The rest of this paper is organized as follows. Section II describes the proposed algorithm. The experimental results and the conclusion are presented in Section III and IV, respectively.

II. Proposed Algorithm

This paper proposed a new algorithm for the overlaid text detection and recognition in a sequence of video frames. There are several problems in the text extraction process e.g. unwanted region, low quality frames, inappropriate threshold for the edge detection in a variety of intensity, similar text color compared to background, improper size of bounding boxes causing over segmented and under segmented results, etc. Therefore, this paper proposed a multiresolution overlaid text detection to detect the candidate text regions in various illuminations and complex backgrounds. The overview of the proposed algorithm and the example of overlaid text are presented in Fig. 1 and Fig. 2, respectively.

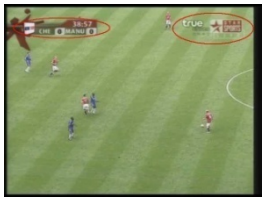


Fig. 2. Example of overlaid text.

A. Adaptive Threshold

In a variety of lighting conditions, a suitable threshold is required to adapt to the changes of intensity throughout the video frames. With a fixed threshold, it may provide a good result in a predefined case. However, if the brightness of the scene changes, the specified threshold might not be appropriate. The high threshold value will result in the low quality edge, while the low threshold value will produce high quality edge with large amount of noise. In this paper, the average intensity (AI) of an image with the size of m x n is adopted to estimate the suitable threshold for edge detection process. The proposed threshold will be adapted to variations of illumination throughout a sequence of the video frames. The proposed adaptive threshold (AT) is presented in (2)

$$AverageIntensity = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m} image(i,j) \tag{1}$$

$$AT = \begin{cases} 0.025 & ;AI < 116 \\ 0.025 - \left[\dfrac{(AI-116)\times 0.007}{31}\right] & ;116 \le AI \le 147 \\ 0.018 & ;AI > 147 \end{cases} \tag{2}$$

B. Background Subtraction

Most problems of overlay text detection are due to the complication of the video frames caused by the overlapping of the desired text and stadium or advertisement. This can cause the difficulties for text detection since it may result in many unwanted region and lead to misunderstanding of the system. This paper proposed the text detection based on background subtraction to reject noise in complex background in order to increase the accuracy rate in text localization process. To construct a background, the average intensity of all pixels in m images is estimated, as in (3). The resulted background is shown in Fig.3. Only overlaid text e.g. time, score, team and channel etc. will remain, the other components e.g. player, football, stadium and viewer etc. will be blur out.

$$Background = \frac{1}{m}\sum_{i=1}^{m} image(i) \tag{3}$$

Fig. 2. The result image obtained from background subtraction process.

C. Multiresolution Generation

Due to low quality of the video frames, most of them contain extensive noises. In order to obtain the efficient results for text detection, a multiresolution approach is an alternative choice, which not only be useful for noise suppression but also improve computational efficiency by allowing signal analysis from coarse-to-fine details. Multiresolution or pyramidal image representation decomposes an image into a number of levels, each of which contains the image at a different resolution. Usually, each level of the pyramid is half of the size of its predecessor as shown in Fig. 4.

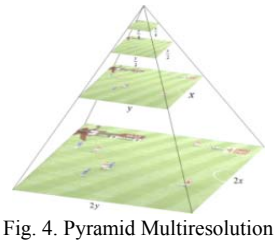


Fig. 4. Pyramid Multiresolution

D. Text Detection

The target of the text detection in video frame is to detect only overlaid text and to eliminate noise or false positive candidate text area. The proposed text detection consists of 3 main processes: Edge detection, Edge map Smoothing and Noise Elimination based on Morphological Opening.

- Edge Detection: Edges provide an opportunity to detect various kinds of the overlaid text with different fronts, colours, sizes and shapes. Edge detection technique can detect strong edges by considering the value of the magnitude gradient (G), which is calculated from the gradient level in horizontal and vertical directions (Gx, Gy) from Sobel edge detection [3]. Then, edge map is generated using proposed adaptive thresholding (2) of the magnitude gradient as shown in Fig. 5(a).

$$|G| = |Gx| + |Gy| \tag{4}$$

$$Edge = \begin{cases} 1 & ;|G| \geq Threshold \\ 0 & ;|G| < Threshold \end{cases} \tag{5}$$

- Edge map Smoothing: Results from edge map are thin edge and may lead to incomplete shape of desired text area. Therefore, the edges should be expanded to allow connections between unconnected edges. Thus, the edge map should be smoothed using morphological dilation technique with the structure element size of 1x7 to enhance the quality of the text region.
- Noise Elimination based on Morphological Opening: The connected edges are formed as candidate text cluster. The morphological opening with the size of 3x23 is used to discard unwanted region and smooth the shape of the candidate text areas. The result from this process is shown in Fig. 5(c).

(a) (b) (c)

Fig. 5. (a) Sobel edge detection. (b) Result from edge map Smoothing.(c)Result from noise elimination based on morphological opening.

E. False Positive Eliminate

The false positive obtained from previous process still remains. In this paper, the geometrical property of the width and height of the text shape is used to eliminate the false positive as shown in (6) and (7).

$$10 \leq Height \leq 55 \tag{6}$$

$$3 \leq \frac{Width}{Height} \leq 14 \tag{7}$$

If the geometrical property of any candidate text block does not follow these rules, it will be considered as a false positive, thus will be discarded. Otherwise, it is accepted as an actual text box, as shown in Fig. 6.

Fig. 6. Bounding boxes: Before and after false positive elimination

F. Text Recognition

In this paper, the candidate text regions are recognized using commercial OCR software (ABBYY FineReader 10). Even though most of the OCR engines claimed to achieve high recognition performance, from experiments, they refused to interpret the overlaid text directly on the text regions of the grayscale image or was not able to localize and segment characters. This means that applying the OCR directly on the image leads to very poor recognition rates. Therefore, efficient detection and recognition of text characters from background is necessary to enhance the performance of the OCR system. Therefore, text enhancement and recognition is proposed to improve the performance of the OCR software.

- Text enhancement is a process to remove the background surrounding text characters and segmented them to form the input to the OCR. Firstly, the candidate text regions obtained from previous process are enlarged to increase the gaps between characters. Then, the adaptive thresholding based on average intensity as presented in (2) will be applied on the enlarged text areas to generate binarization of the candidate texts. After that, the binarized text is partitioned into characters based on connected component analysis. Finally, the segmented characters are renormalized to the actual resolution and sent to the OCR.
- Text Recognition: In this process, the segmented characters are recognized using the OCR software namely ABBYY FineReader 10. This software requires the format of the characters to be black inside character region and white for the background. However, the recognition rate depends on the quality of the segmented characters. If the quality of the segmented characters is not good, the OCR may fail to identify the correct answer as shown in Fig. 7. For example, with the low quality of the letter "E", it maybe misleads to the letter "t" in some cases.

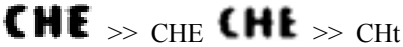


Fig. 7. Example of good and bad recognition

G. Keyword Matching for Character Refinement (KMCR)

The performances of OCR systems closely rely on the quality of the targeting character regions, which will drop abruptly due to the low quality or poor resolution of the characters. In order to overcome this drawback, the keyword matching based on character refinement (KMCR) is proposed to search on the soccer concept for each video frame and the search results are temporally collected and

voted for the final best matches at the final. Instead of performing a simply search to match for a single keyword in each frame as in [1]. In the proposed searching process, the interpolation search algorithm has been adopted to search for the identified words from the indexed database e.g. team name, player name, time, score, and important events, etc. The identified words are used to match to the target recognition results and the confident score will be calculated for each result. If no match has been found, the search will reduce to the part before and after the estimated words. Then, the decision will be made based on the weighting scores of the estimated words. The matched keywords for each frame are temporally tracked throughout ten consecutive video frames. Then, the confident scores will be evaluated to obtain the best match for each case. The example of the results from the proposed KMCR algorithm is presented in Table I, where the incorrect recognition words "CIIE", "CHt", and "(HE" matched to the team name "CHE". The KMCR helps to increase the accuracy and reliable to the recognition results, thus leads to an improvement on the video indexing.

TABLE I: EXAMPLE OF KMCR ALGORITHM

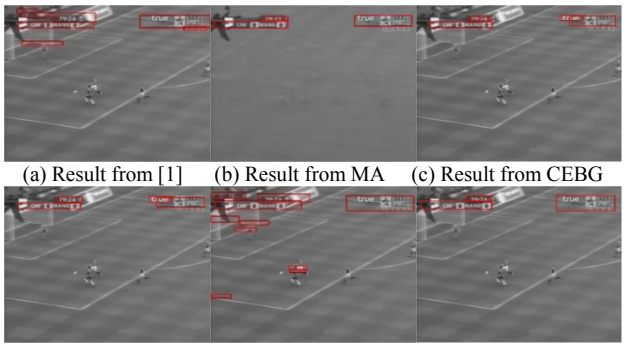
Example	Recognition Result	KMCR Results
1	CIIE	CHE
2	CHt	CHE
3	(HE	CHE

III. EXPERIMENTAL RESULT

There has not yet been a standard video frame for overlaid text detection. This paper uses video soccer F.A. Cup 2007 Final Chelsea vs. ManU in the experiments. The resolution and frame rate of the test video sequence are 640 x 480 pixels, 30 frames per second respectively. The test video sequence is partitioned into 3 test sets based on illumination distribution in the video frames, dark, light, combination. The proposed algorithms were implemented in MATLAB programming language on an Intel(R) Core(TM)2 Duo 2.00 GHz CPU 2.0 GB DDRIII Ram.

Fig. 8. Data set (1) Dark (2) Dark and light (3) Light

In the experiments, the performances of six text detection techniques are compared. The first technique is adopted from [1]. In the second technique, the candidate text regions are defined using background subtraction based on multi-frame averaging (MA). Integration of background subtraction to the edge detection process is adopted in the third technique (Combined Edge detection and Background subtraction for text detection: CEBG). The results of edge detection will be refined by the results of the background subtraction to reduce the false positive regions. The forth technique performs (Noise Elimination for CEBG: NCEBG) noise elimination based on morphological operation on the candidate text regions obtained from CEBG to eliminate the unwanted regions and smooth the shapes of the candidate text regions. The multiresolution text detection in [4] is adopted and evaluated in the fifth techniques. The last text detection technique is the proposed multiresolution approach with adaptive thresholding, background subtraction, and noise elimination.



(a) Result from [1] (b) Result from MA (c) Result from CEBG

(d) Result from NCEB (e) Result from [2] (f) Result from proposed algorithm

Fig. 9. Text detection result obtained from various algorithms

The text detection performance is evaluated in the form of precision (P) and recall (R) rate as in (8) and (9).

$$P = \sum_{i=1}^{N} \left(\frac{D_i \cap GT}{D_i} \right) \tag{8}$$

$$R = \sum_{i=1}^{N} \left(\frac{D_i \cap GT}{GT} \right) \tag{9}$$

D_i is detected text bounding box area in frame i, GT is ground truth text bounding box area from manual cut, and N is number of the detected bounding boxes. The performance comparison of those six text detection techniques is presented in Table II.

TABLE II: RESULT OF TEXT DETECTION

	[1]	MA	CEBG	NCEBG	[2]	Proposed
# of GT Bounding boxes	45	45	45	45	45	45
Precision (%)	73.54	76.89	86.62	81.20	61.17	72.16
Recall (%)	82.68	81.22	56.75	73.30	91.16	89.40

From the experimental results, the text detection algorithm from [1] generates over-segmented bounding boxes with several false alarms (Fig. 11 (a)) while MA seems to better eliminate noises in the complex background but still creates longer bounding boxes (Fig. 11 (b)). CEBG can achieve high precision but very low recall rate, which is similar to NCEBG. Even though the text detection algorithm from [2] can achieve very high recall rate, it produces numerous false alarms with over-segmented bounding boxes as shown in Fig. 11 (e). Since the proposed algorithm is developed based on multiresolution approach similar to [2] thus combines noise reduction based on morphological opening similar to NCEBG, the experimental results show that the proposed algorithm can better reject most of the false alarms compared

to [2] while can achieve higher recall rate compared to NCEBG. This proves that the proposed algorithm can reasonably compromise between precision and recall.

After the text bounding boxes have been extracted and performed quality enhancement, they are segmented into character regions. The segmented character regions after enhancement can achieve higher precision and recall. These characters are then recognized using the OCR software. We evaluate the performance of the text recognition via Character Accuracy Rate (CAR) as in (10) and Character Error Rate (CER) as in (11):

$$CAR = \frac{TR}{GT} \tag{10}$$

$$CER = \frac{FR}{GT} \tag{11}$$

TR is true recognition and FR is false recognition in ground truth.

Three text enhancement techniques: Noise Reduction using Erosion (NRE), Noise Reduction using Opening (NRO) and proposed text enhancement are developed for recognition performance comparison. From table III, the proposed text enhancement provides 63.88% CAR with low miss recognition. From the results, we can see that even though the proposed text enhancement can better achieve higher accuracy rate, the CAR is still low due to the low quality of the video frames. The proposed KMCR algorithm can be integrated to perform keyword searching and temporally tracking in order to increase text recognition performance.

TABLE III: RESULT OF TEXT RECOGNITION

Algorithm	CAR	CER	Miss
NRE	54.12	17.17	28.71
NRO	30.48	16.93	52.59
Proposed	63.88	19.82	16.30

IV. CONVLUSIONS

This paper proposed multiresolution overlaid text detection and recognition. The text detection algorithm is performed based on the proposed adaptive threshold and integrated to the background subtraction to detect the candidate text regions at the top level of the pyramid. The multiresolution detection allows us to reduce noise and unwanted regions in the video frame while the proposed adaptive threshold helps to automatically select a suitable threshold for various illumination conditions. The candidate text regions are then enhanced and segmented into character regions using the proposed text enhancement algorithm. The segmented characters are recognized using optical character recognition. Due to the low quality of the video frame, the OCR results in poor character recognition rate. However, with the proposed KMCR, it helps to increase the accuracy and reliability to the recognition results, thus leads to an improvement on the video indexing.

REFERENCE

[1] A. Halin, M. Rajeswari, and D. Ramachandram, "Overlaid Text Recognition for Matching Soccer-Concept Keyword," *Fifth International Conference on Computer Graphics*, Imaging and Visualisation, pp. 235-241, 2008.

[2] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "Multiresolution Text Detection in Video Frames," *VISAPP 2007 - International Conference on Computer Vision Theory and Applications*, pp. 161-166, 2007.

[3] R. Gonzalez and R. Woods, "Digital Image Processing. 3rd ed," *Pearson Education*, Inc, Upper Saddle River, New Jersey 07458, pp. 649-710, 730-732, 2010.

[4] V. Kobla, D. Dementhon, and D. Doermann, "Identifying sports videos using replay, text, and camera motion features," In *Proc SPIE*, pp 332-343, 2000.

[5] T. H. Tsai. Y. C. Chen, and C. L. Fang, "A Comprehensive Motion Videotext Detection Localization and Extraction Method," *IEEE 23rd International Conference on Data Engineering Workshop*, pp. 515-519, 2007.

[6] T. Phan, P. Shivakumara, and C. Tam, "A Laplacian Method for Video Text Detection," *10th International Conference on Document Analysis and Recognition*, pp 66-70, 2009.

[7] S. Vivier, G. Bossco, and C. Nguyen, "Multiresolution approach for image processing," *Technical report*, Eransmus ICP-A-2007, Leiden, Pays-Bas, April 1996.