Instantaneously Responsive Subtitle Localization and Classification for TV Applications

Bahman Zafarifar[1,2], Jingyue Cao[2] and Peter H. N. de With[2,3], *IEEE Fellow*
[1] Trident Microsystems, [2] Eindhoven University of Technology, [3]CycloMedia Technology

***Abstract*—This paper presents an algorithm for localization and classification of subtitles in TV videos. We extend an existing static-region detector with object-based adaptive filtering and binary classification of subtitle bounding boxes, using geometry and text-stroke alignment features. Compared to this static-region detector, we reduce the number of falsely detected subtitle pixels by a factor of 20, at the cost of only 2% of subtitle pixels.**

I. INTRODUCTION

Extensive research has been performed on subtitle/text detection and recognition, using analysis techniques such as connected-components [2], Texture-based [3], Edge-based [4] and multiple-frame information integration [5]. Nevertheless, *reliable* detection of subtitles overlaid on top of the diverse TV video content remains challenging.

In modern high-end TV sets, specific circuitry/software enables detection and special treatment of static regions like logos and subtitles. We present an algorithm that builds upon the static-region detector of a commercially available TV video processing chip [1], taking it from a pixel-based still/non-still decision to an *object-based* subtitle/non-subtitle decision. The algorithm responds to (dis)appearing subtitles instantaneously, requires access to only one video frame at a time, and provides two types of information: (1) a set of bounding boxes around blobs of static regions, categorized by their geometry and filling degree, and (2) bounding-box level text classification using text-stroke alignment features.

We describe the algorithm in Section II and provide experimental results and conclude in Section III.

II. ALGORITHM DESCRIPTION

A. Refined static-region detection

Fig.1 shows the overview of the algorithm. Our starting point is an existing static-region detector [1] that uses pixel-based motion and contrast information. This static-region detector finds nearly all subtitle areas, but also accepts many other static edges (Fig. 1(b)). We prune its results as follows. We identify steep horizontal luminance transitions and denote a pair of these transitions (rising-followed-by-falling or vice versa) that are within a small horizontal distance as a *transition pair*. The static-region map is now pruned by retaining only areas that have a high density of transition pairs. Fig. 2 (b) and (c) compare the transition pairs of non-subtitle and subtitle areas, and Fig. 1 (c) shows the pruned static-region map.

This work has been sponsored by NXP Semiconductors and the Dutch NWO Casimir program.

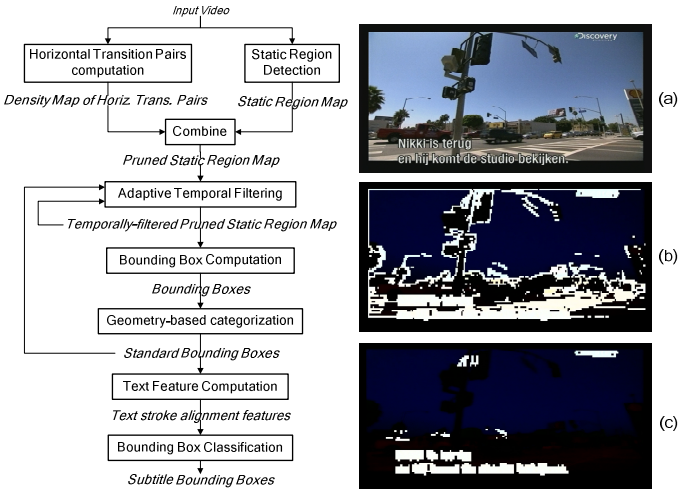

Fig.1. left: Algorithm overview, (a) Input image, (b) Static-region map, (c) Pruned static-region map.

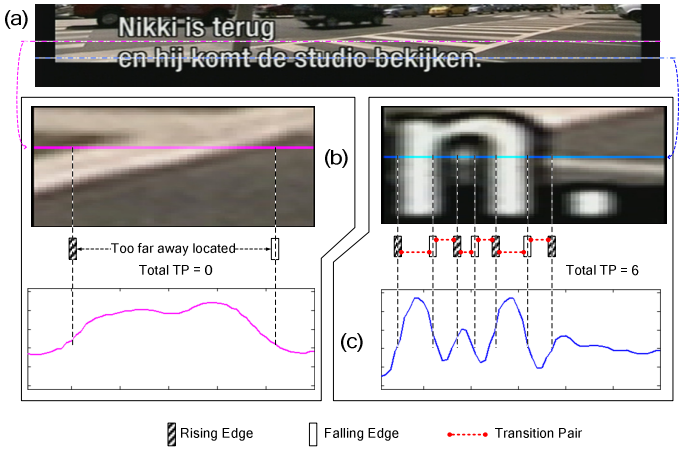

Fig.2. (a) Input image, (b) Horizontal transition pairs of a non-subtitle region, (c) Horizontal transition pairs of a subtitle region.

B. Object-based adaptive temporal filtering

We devise a strategy for content-adaptive filtering of the pruned static-region map, which improves the subtitle detection stability, while yielding *instantaneous* response to (dis)appearing subtitles. This is achieved by feeding back the object-level subtitles bounding box information to the pixel-level filtering operation, as follows. Let α be the temporal filter coefficient, where a high value of α means quick adaptation to current frame's pruned static region map. *Outside* slightly expanded bounding boxes (computed on previous frame), α is set to maximum. This yields a quick response at appearing subtitles. *Within* each expanded bounding box, α is proportional to the amount of difference between the current frame's pruned static-region map and its

temporally filtered version from the previous frame. This yields a quick response at disappearing subtitles while performing strong filtering at sustaining ones, since at disappearing subtitles the current frame's pruned static region map differs significantly from previous frames'.

C. *Bounding-box computation and categorization*

In this step, we compute bounding boxes around blobs of static regions. First, we compute initial bounding boxes by applying a fixed threshold to bi-projections of the temporally filtered pruned static-region map. In a next refinement step, each initial bounding box goes through iterative bi-projections that use an adaptive threshold, until the bounding box can no longer be split (Fig. 3). Finally, we categorize bounding boxes to standard or non-standard based on their geometry (height and width) and the filling degree of pruned static-region map.

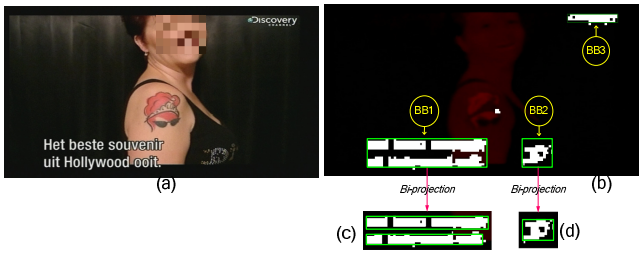

Fig. 3. Bounding box computation. (a) Input image, (b) Pruned static-region map and initial bounding boxes: BB1 contains two subtitle lines and BB2 is larger than the inscribed static regions, (c),(d) Refinement of BB1 and BB2: the two subtitle lines inside BB1 are separated and the size of BB2 is reduced.

D. *Classification based on text-stroke alignment*

We have observed that the number of vertical and horizontal text strokes have a certain ratio, and use this text-specific feature to classify standard bounding boxes to (non-)subtitle classes. For each bounding box, we compute a vertical text-stroke alignment feature (Fig. 4 (b)) by taking the average of the convolution of a 9×3 box filter with a map that contains a '1' at the center of a horizontal up-down transition pair and is '0' elsewhere. A horizontal text-stroke alignment feature (Fig. 4 (c)) is computed similarly on the transposed image.

Using a manually annotated training set, we estimate the parameters of a Gaussian model (Fig. 4 (d)) for the feature distribution of the subtitle class. The model is then employed for binary classification of each standard bounding box to (non-)subtitle classes, based on the mentioned two features. Fig. 4 (e) shows the classifier's performance on our test set.

III. RESULTS AND CONCLUSION

For training and testing the classifier, we used two video sets, each more than 5000 frames. With the test set, we obtain a bounding-box classification result (Table 1), which indicates that 99.66% (TP/(TP+FN)) of subtitle bounding boxes are correctly classified, while 92.82% (TN/(TN+FP)) of non-subtitle bounding boxes are correctly rejected. This shows the distinctive properties of the text-stroke alignment features.

We have also performed a pixel-level evaluation as follows. We create a *pixel-level ground-truth* map using the static-region detector, by assigning '1's to static pixels lying within manually annotated subtitle bounding boxes, and '0's elsewhere. This map is compared to a *pixel-level subtitle-detection map* created by assigning '1's to static pixels that lye within computed subtitle bounding boxes, and '0's elsewhere. Table 2 shows the pixel-level evaluation results. An improvement of factor of 20.6 (FP_{srd}/FP_{prop}) is obtained in rejecting non-subtitle pixels, while loosing only 2% (1-TP_{prop}/TP_{srd}) in the correct subtitle pixels.

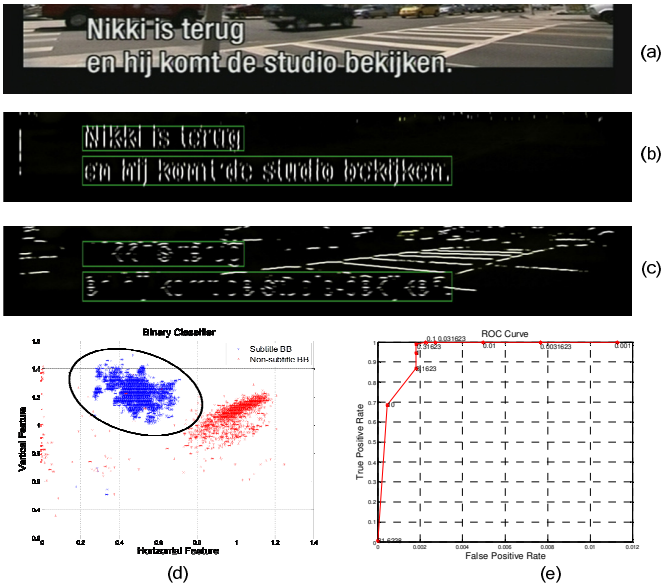

Fig. 4. Classification based on text-stroke alignment features. (a) Input, (b),(c) Vertical and horizontal text features, respectively, (d) Feature distribution of the training set, (e) ROC curve of classification result on the test set.

We conclude that the proposed algorithm offers high-level information about the existence and location of subtitles, with instantaneous response. The algorithm may enable accurate segmentation of static text regions, which can be used for prevention of artifacts at subtitle regions and their surrounding areas in TV motion-compensated frame-rate up conversion.

TABLE 1 – EVALUATION RESULTS OF SUBTITLE CLASSIFICATION

	True Pos.	False Neg.	True Neg.	False Pos.
BB No.	6691	23	2793	216

TABLE 2 – PIXEL-LEVEL EVALUATION RESULTS

	TP	FP	TN	FN
Static-Region Detector	2021314	1585637	69113049	0
Proposed method	1980696	76945	70621741	40618

IV. REFERENCES

[1] PNX5100 Nexperia video back-end processor, www.nxp.com.

[2] Soo-Chang Pei and Yu-Ting Chuang, "Automatic text detection using multi-layer color quantization in complex color images," *Int. Conf. on Multimedia & Expo*, 2004, pp. 619-622.

[3] K.I. Kim, K. Jung, and J.H. Kim, "Texture-Based Approach for Text Detection in Images Using Support Vector Machines and Continuously Adaptive Mean Shift Algorithm," *IEEE Trans. on Pattern Anal. And Mach. Intell.*, vol. 25, 2003, pp. 1631-1639.

[4] H. Li, D. Doermann, and O. Kia, "Automatic text detection and tracking in digital video," *IEEE Trans. on Img. Proc.*, vol. 9, 2000, pp. 156, 147.

[5] X. Tang, X. Gao, J. Liu, and H. Zhang, "A spatial-temporal approach for video caption detection and recognition," *IEEE Trans. on Neural Networks*, vol. 13, 2002, pp. 961-971.