Real-time Caption Detection and Localization in TV Programs via Stroke Width Transform and Morphological Transform

Jianghua Cheng*, Gui Gao,Xishu Ku, Jurong Liu and Yongfeng Guan
College of Electronic Science and Engineering, National University of Defense Technology, Changsha, 410073, China
*jianghua_cheng@nudt.edu.cn

ABSTRACT
Captions in TV programs are important clues for video content analysis, indexing, and retrieval. However, it is difficult to extract due to complex background, unknown caption color, etc. Nevertheless, caption detection and localization on its own is a hard task, let alone real time realization. In this paper, we propose a novel algorithm to extract caption from TV channel videos. Our proposed algorithm is composed of three steps: canny edge detection and binarization, simple stroke width transform, and histogram projection analysis. The experimental results show that the proposed method mentioned above is effective in real time TV caption detection and localization.

KEYWORDS
real-time, caption, detection, localization, morphological transform, stroke width transform.

1 Introduction

Captions in TV programs carry rich and useful information, which can provide TV viewers with better visual understanding. Currently, most TV stations have used the overlay captions in broadcasting videos to convey more semantics and deliver better viewing experience. For example, captions in news always indicate where, when, and who of the highlight events, to which the audiences usually pay more attention.

With the rapid growth of TV multimedia data, there is an urgent demand for video semantic understanding operations like analysis, retrieval and annotation. Caption information is buried in videos, and is highly compact and summarized. It can be used for video understanding. The extraction of caption information has attracted the researchers' particular attention in the world over the past decades. A large number of techniques have been proposed to address this problem. The procedure of caption information extraction from TV programs is usually divided into four steps: detection, localization, extraction, and recognition. Detection and localization are the first two steps among these four ones, performance of which would influence the result of two latter ones.

The detection step roughly distinguishes caption regions from background. The localization step determines the accurate boundaries of caption strings by refining the detected caption regions. Up to present, many great achievements in caption detection and localization have been made by researchers. Among them, representative features such as color, edge, and texture-based information, are frequently used to distinguish caption regions from others. The color-based approaches assume that caption strings are usually composed of a uniform color. Kim et al. [1] believed that the color of caption is spatio-temporally consistent, which can be used to calculate color matching score for detecting caption regions with the help of orientation consistency. Shivakumara et al. [2] applied wavelet decomposition on three color bands separately to increase the gap between text and non text pixels. Additionally, similar method like Connected Component Analysis (CCA) is frequently applied to detect homogeneously colored connected caption areas [3, 4]. However, it is rarely true that the overlay text

consists of a uniform color due to degradation resulting from compression coding and low contrast between text and background [5]. The edge-based methods consider that caption regions contain rich edge information. Michael et al. [6] proposed a local thresholding method on a sobel-based edge strength map. Chen et al. [7] extracted texts by detecting horizontal and vertical edges with a canny operator. These type methods perform well only if there is no complex background. Actually, this condition can't be met because TV video image often contains large numbers of edges in the background. The texture-based methods consider the appearance of text as a special texture to discriminate text from non-text, which usually divide a whole image into blocks to calculate the texture features of blocks. Li et al. [8] used the mean, second, and third-order central moments in the wavelet domain as texture features. Crandall et al. [9] used DCT coefficients of intensity images as texture features. After calculating the texture features, neural network or support vector machine were employed to classify text blocks and non-text blocks. The main shortcoming of the methods attributed to this category is the high computational complexity.

Although many methods have been proposed for this task in the last decade, this problem is still challenging due to the complex background, the non-uniform illumination, the variations of text font and size. To the best of our knowledge, few works have been done to address real-time caption detection and localization especially.

This paper proposes a new method that emphasizes on the real-time achieving caption detection and localization. The rest of this paper is organized as follows. In Section Ⅱ, a general scheme of real-time caption detection and localization is introduced, and simple stroke width transform is proposed. In Section Ⅲ, experiments are described and the performances of caption detection and localization are discussed. In Section Ⅳ, some conclusions are reached.

2 Methodology

2.1 Characteristics of captions

Before introducing the proposed algorithm, we sum up caption characteristics that are frequently used in video caption detection and localization on the basis of observation and review of previous methods [6, 10, 11].

- Contrast: The contrast is usually high between caption regions and background. However, this cannot be guaranteed absolutely due to complex background in video scenes.
- Color: Typically, characters in a single subtitle have identical color, but not all captions have the same color. Additionally, the lossy compression causes the color bleeding effect at the edge of caption.
- Orientation: Caption characters are generally aligned horizontally and appear in clusters. Thus, the constraint of horizontal caption orientation is adequate for the video understanding purpose.
- Location: Most stationary captions remain on the screen for a minimum duration of several seconds. Scrolling captions are used in special effects report, which move in a horizontal way.
- Font Size: To be easily readable at normal viewing distance, caption font should have a minimum size. Caption font should also have an upper-bounded size because captions should not shelter the broadcasting TV video image.
- Stroke Density: A caption containing region usually possesses high stroke density. These stroke-like structures are important clues and distinct features for character detection.

Based on the analysis of TV caption characteristics, we can regard caption characters as a special type of symbol that has the union of the above characteristics.

2.2 Framework

The framework overview is illustrated in Figure 1. Real time input analog TV video signal is sampled by a PCI video capture card. Then each frame is captured and converted into a 256-level grayscale image, which becomes the current frame. We do not use the color characteristic, since caption color may be influenced by the background due to limited spatial resolution, or degraded by analog/digital sampling noise. After current frame capture, canny edge operator is applied. The canny operator coarsely detects edges of the current frame and binaries it. All these three edge detection results are feed into stroke width transform step to detect strokes of characters for robust caption localization. Morphologic operations like open and close are used to merge disconnected strokes into regions and remove isolated noisy edges. Based on region area computing, small regions are removed. Finally, histogram projection analysis like horizontal and vertical projection analysis is used to determine the boundary of the caption blocks, and colored bounding boxes are drawn on the caption regions.

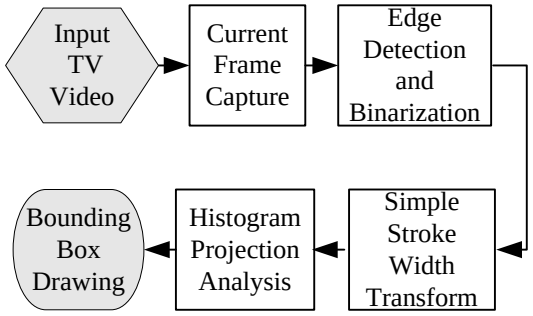


Figure 1. Flowchart of the proposed method.

2.3 Edge detection and binarization

"Dense edges" is one of the most prominent characteristics of caption characters, which can be employed to find possible caption regions.

Canny operator [12] is used to detect caption edges. Let $I(x,y)$ denote current frame, and $G(x,y)$ is the 2-D Gaussian function. We have the following frame edge intensity and orientation formulation respectively:

$$C(x,y) = |\nabla G \otimes I(x,y)| \quad (1)$$

$$\vec{n} = \nabla G \otimes I(x,y) / |\nabla G \otimes I(x,y)| \quad (2)$$

where ∇G is the gradient of G, and $\otimes$ stands for the convolution operation.

Then, Otsu's thresholding is used to determine the threshold for edge binarization. Supposing that the edge intensity has *L* levels, there are *M-1* thresholds $\boldsymbol{T} = \{T_1 \cdots, T_k \cdots, T_{M-1}\}$, which divide the edge intensity into *M* classes. The optimal thresholds are chosen by maximizing the between-class variance as:

$$\boldsymbol{T}^* = \arg \max_{1 \le T_1 < \cdots < T_{M-1} < L} \{\sigma_B^2(T_1 \cdots, T_k \cdots, T_{M-1})\} \quad (3)$$

where $\sigma_B^2 = \sum_{i=1}^{M} \omega_i (\mu_i - \mu_T)^2$ is the between-class variance, with ω_i and μ_i being the zeroth-order and the first-order cumulative moments of the histogram up to the *i*th class, respectively, and μ_T being the total mean level of $C(x,y)$.

2.4 Simple Stroke Width Transform

As a basic element of caption characters, strokes provide robust features for caption detection in TV images. Caption can be modeled as a combination of stroke components with varies of orientations. The constant stroke feature like stroke width can be utilized to effectively remove the interference of non-stroke edges in complex background so as to make the detection and localization of captions much more accurate [13].

The Stroke Width Transform (SWT) is a local image operator which computes per pixel the width of most likely stroke containing the pixel [14]. However, the traditional SWT needs to calculate the current edge point's corresponding symmetrical edge points throughout the whole image, which consumes lots of time, and plenty of false edge points would be detected.

We propose a modified SWT (Figure 2) that only calculates the symmetrical edge points in the neighboring area, based on the fact that the caption's stroke width of dominating language characters wouldn't be large enough. Therefore, fast detection speed is achieved. After that, morphological operations like open, close,

dilate are used to remove the isolated edges and connect the separated caption edges into blocks.

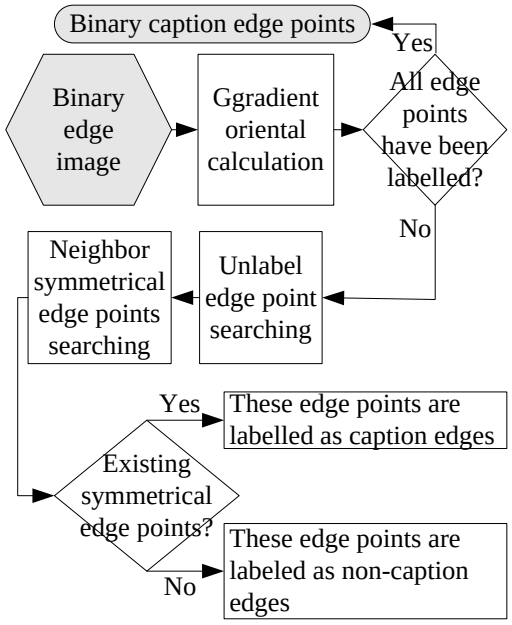


Figure 2. Simple SWF caption edge points detection flow chart.

2.5 Histogram Projection Analysis

There are plenty of detected true blocks that include multi-line captions, and false blocks that have other objects. In order to distinguish the true caption lines from the false and separate these blocks into single-line caption, we use the histogram projection profile [15]. A horizontal/vertical histogram projection is defined as the sums of the pixel intensities over each column/row. The horizontal and vertical projections are defined as follows.

$$HP(y) = \sum_{x_1}^{x_2} SWT(x, y) \tag{4}$$

$$VP(x) = \sum_{y_1}^{y_2} SWT(x, y) \tag{5}$$

where *SWT(x,y)* is the simple SWT results in Section 2.4. If *HP(y)* is greater than a certain threshold, row *y* is potentially a part of caption line. Similar results are arrived to *VP(x)*. Finally, different words on the same caption line are merged if they are close to each other.

3 Experimental results and analysis

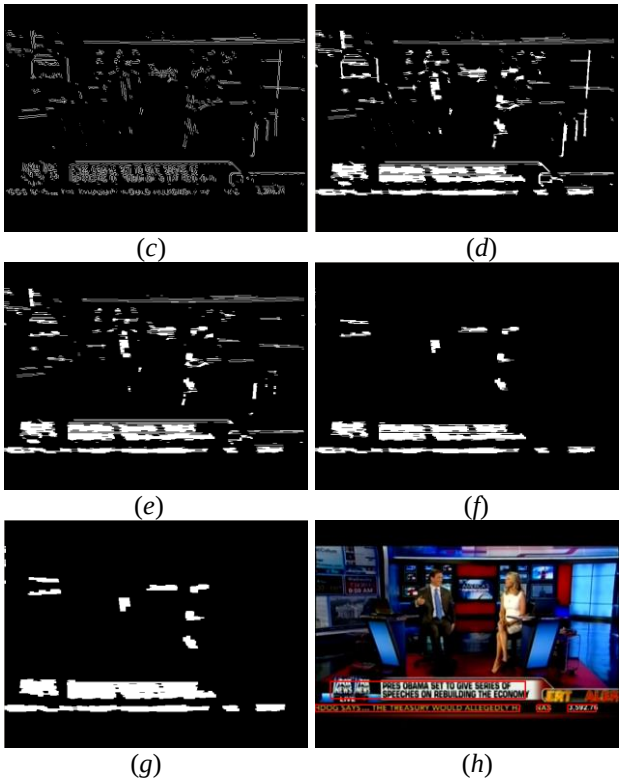
In order to test the performance of the algorithm, about 5 video sequences with 640*480 resolution containing multi-language captions in different font sizes are used. Those videos are recorded into MPEG-2 coded formats. Different video segments are used to test the performance of the proposed algorithm, i.e. Foxes news, movies, and live sports, which last about 60 min. All the experiments are run on P(R) dual-core 2.8GHz CPU, with 2GB SDRAM, and the software platform is VS2008+Opencv 2.4.3. Table 1 shows the mean result with parameters $th1 = 70$, $th2 = 255$ (low and high thresholds of canny binarization), $swidth = 5$ (stroke width), and mint=20 (interval of horizontal neighboring bounding box). From Table 1, we can see that our proposed algorithm can detect and localize captions in real time, and the recall and precision rates are both satisfactory.

Table 1. Mean result of different channels experiment.

No.	Channel	Speed (sec/frame)	Recall Rate (%)	Precision Rate (%)
1	Fox News	0.31	89	95
2	Channel News Asia	0.29	92	97
3	Fox Sport	0.33	96	96
4	Formosa	0.25	95	98
5	Bloomberg	0.36	87	91

In addition, we select a frame from Fox news channel, and present details of each step introduced in Section 2. Fig.3 shows the performance of each operation. Fig.4 shows other representative channel frame experimental results.

(*a*) (*b*)



(*c*) (*d*)

(*e*) (*f*)

(*g*) (*h*)

Figure 3. Each step experimental result. (a) Original captured frame. (b) Canny edge detection and binarization. (c) Simple SWT. (d) Morphological close operation with rectangle structure element (5,1). (e) Morphological open operation with rectangle structure element (5,1). (f) Contours calculation and small area remove. (g) Morphological dilate operation with rectangle structure element (5,1). (h) Caption bounding boxes are drawn in red.

(*a*) (*b*) (*c*)

Figure 4. Random selection TV channel frames experimental result. (a) Channel News Asia. (b) Formosa. (c) Bloomberg.

4 Conclusions

This paper proposes a real time caption detection and localization method based on simple stroke width transform and morphological transform. The experimental results show that the algorithm can detect and localize TV captions fast and accurately. In addition, this method is robust to various font sizes, font styles, contrast levels, and background complexities.

REFERENCES

[1] D. Kim, and K. Sohn, "Static text region detection in video sequences using color and orientation," 19th International Conference on Pattern Recognition, 2008, pp.1-4.

[2] P. Shivakumara, T. Q. Phan, and C. L. Tan, "New wavelet and color features for text detection in video," International Conference on Pattern Recognition, 2010, pp.3996-3999.

[3] X. Chen, J. Yang, and J. Zhang, A. Waibel, "Automatic detection and recognition of signs from natural scenes," IEEE Transactions on Image Processing, vol.13, 2004, pp.87–99.

[4] R. Jiang, F. Qi, L. Xu, and G. Wu, "Using connected-components' features to detect and segment text," Journal of Image and Graphics, vol.11, 2006, pp.1653-1656.

[5] W. Kim, and C. Kim, "A new approach for overlay text detection and extraction from complex video scene," IEEE Transactions on Image Processing, vol.18, 2009, pp.401-411.

[6] 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, 2005,pp.243-255.

[7] D. T. Chen, J. M. Odobez, and H. Bourlard, "Text detection and recognition in images and video frames," Pattern Recognition, vol.37, 2004, pp.595–608.

[8] H. Li, D. Doremann, and O. Kia, "Automatic text detection and tracking in digital video," IEEE Transactions on Image Processing, vol.9, 2000, pp. 147–156.

[9] D. Crandall, S. Antani, and R. Kasturi, "Extraction of special effects caption text events from digital video," International Journal on Document Analysis and Recognition, vol.5, 2003, pp.138–157.

[10] B. Zafarifar, J. Y. Cao, and H. N. Peter, "Instantaneously responsive subtitle localization and classification for TV applications," IEEE Transactions on Consumer Electronics, vol.57, 2011, pp.274-282.

[11] K. Jung, and K. I. Kim, "Text information extraction in images and video: a survey," Pattern Recognition, vol.37, 2004, pp.977-997.

[12] J.Canny, "A computational approach to edge detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.8, 1986,pp. 679-698.

[13] X. Liu, and W. Wang,"Robustly extracting captions in videos based on stroke-like edges and spatio-temporal analysis," IEEE Transactions on Multimedia, vol.14, 2012, pp.482-489.

[14] B. Epshtein, E.Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp.2963–2970.

[15] X. Gao, and X. Tang, "Automatic news video caption extraction and recognition," 2nd Int. Conf.

Intell. Data Eng. Automated Learning Data Mining, 2000, pp. 425–430.