A Method of Caption Detection in News Video

He HUANG, Ping SHI[1]

Abstract. News video is one of the most important media for people to get information. However, it is an urgent problem to find the useful information from a huge amount of news video efficiently and correctly. The caption in news video highly summarizes the related news story and can be used for effective retrieval. In this paper, based on the feature of news video, a method of caption detection in news video is proposed. Firstly, the key frames with captions are detected by using color and edge information. Then, the caption text is extracted by Otsu algorithm. The experiment results show that all the caption frames can be detected and combined with OCR software, the proposed method can give an average recognition rate of 87.3%.

Keywords: news video • caption detection • edge detection • Otsu

1. Introduction

As a typical video, news video is the major path through which people could get informed. However, with the accelerating pace of life and a significant increase in news events, the regularly news video on TV is no longer able to meet people's needs, and the internet is now growing so fast that it becomes the main way for people to watch the news. Although provided the freedom to choose any videos as we like, the traditional linear way to find what we need is time-consuming and low efficient. Moreover, important information is often missed. Therefore, it's

[1]He HUANG (✉)
Information Engineering School, Communication University of China,
Beijing, China
e-mail: hhcucgrow@cuc.edu.cn
Pin SHI(✉)
Information Engineering School, Communication University of China,
Beijing, China
e-mail: shiping@cuc.edu.cn

necessary to establish news video library for content-based news video retrieval to solve this problem and caption detection is the most important part.

The characteristics of news video make the establishment of news video library possible. First, the fixed news video production makes the boundary of each news story clear. Secondly, the title caption of news video providing the basis for video retrieval, since it serves not only as the summary of the news story, but also an important sign of the news video structure.

Many research projects have engaged in detection and recognition of caption in recent years [1]-[6]. Huang at al. [1] performs Harris corner detection on stroke map of detected text lines which is based on Log-Gabor filters. Then morphological operation is utilized to connect these corners into text regions. Zhao at al. [2] uses a corner based approach which is inspired by the observation that there exist dense and orderly presences of corner points in captions. Leon et al.[3] develops a method combining texture and geometric features to detect captions and also takes advantage of the region-based image model. Sharma at al. [4] presents a new method based on dominant text pixel selection, text representatives and region growing arbitrarily-oriented text detection in video. Cai at al. [5] proposes an algorithm based on edge detection, threshold calculation and edge size limitations, filters non-text regions by the scope of the text pixel density. T.Sato at al.[6] first deals with an image with a 3×3 horizontal differential filter, then extracts vertical edge features with a suitable binary threshold, finally gets independent caption region by detecting aggregated regions and calculating the rectangles around.

In this paper, we present an effective method to capture captions in news video: first picking up the key frames with captions by color statistic and edge detection and then obtaining the caption text with Otsu algorithm

2. Analysis of the caption text in news video

The texts in news video can be divided into two categories [7]: scene texts and caption texts. Scene texts are the part of the image captured by the camera, such as texts in scenes and the license plate numbers, as shown in Fig.1. Because of the unfixed location and shown time, it is generally difficult to detect scene texts. Yet caption texts summarize the main information of news story and can be extracted easily because of the fixed location and shown time, as shown in Fig.2. There are some common properties of caption texts in most news video programs, as summarize below:

1. The caption texts in news video have fixed size and fonts in the same news video programs.
2. There is always a rectangle background behind the caption texts.
3. There is a strong contrast between the background color of caption texts and the image frame color.
4. The location of the caption texts in the same news video programs is fixed.

5. The caption texts stay in the screen for at least several seconds. According to rough statistics, they can last 5 seconds to 20 seconds.

Fig. 1 Scene texts

Fig. 2 Caption texts

3. Caption detection

In this section, we detail the process and algorithms of key frame detection and caption text extraction.

3.1 Caption detection process

News video is composed of a sequence of image frames. Therefore, the image frames should be picked up from the news video, and the problem of caption detection in news video is then converted into caption detection in news images. The procedure of caption detection in news video is shown in Fig.3.The texts recognition is implemented in the OCR software, it's not included in this paper.

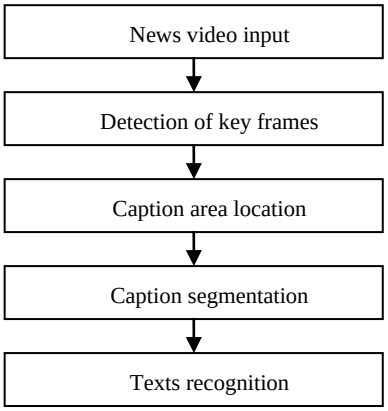

Fig. 3 The procedure of caption detection

3.2 Detection of key frames

This step is to detect the key frames which have captions and the different frames with the same captions should be abandoned. The most obvious characteristic of key frames is that there is caption border after edge detection [8].

Comparing results of several edge detections, we find the texts are clear but the caption border is incomplete after Roberts edge detection, which can be formulated as

$$g(x,y)=\{[f(x,y)-f(x+1,y+1)]^2+[f(x,y+1)-f(x+1,y)]^2\}^{1/2} \quad (1)$$

Where f(x,y) is the input image.

The texts are incomplete but the caption border is clear after Sobel and Prewitt edge detection. Prewitt edge detection operator has two operators which generally known as the template, one is horizontal, the other is vertical, each approaching a partial derivative, as shown in formulae (2).

$$p_v=\begin{bmatrix}-1 & -1 & -1\\ 0 & 0 & 0\\ 1 & 1 & 1\end{bmatrix} \qquad p_h=\begin{bmatrix}1 & 0 & -1\\ 1 & 0 & -1\\ 1 & 0 & -1\end{bmatrix} \quad (2)$$

The difference between Sobel operator and Prewitt operator is that they use different templates. Sobel operator is shown in formulae (3).

$$s_1=\begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix} \qquad s_2=\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix} \quad (3)$$

In conclusion, we choose Prewitt edge to detect caption border and Roberts edge to get frame-to-frame differences.

News videos contain two kinds of captions, one of which is the title caption which contains important desired semantic information, and the other is dialogue in an interview which should be abandoned. Differences of background color in these two kinds of captions could be applied as basis of detection. Therefore, we detect topic caption by statistical color characteristics based on region information.

The title caption remains unchanged in the screen for at least 5 seconds. In order to accelerate operating speed, we detect the key frames at intervals of 5 seconds. This may give rise to redundancy when the duration of a caption is above 10 seconds and the different frames with the same captions are left out by frame-to-frame differences.

The procedure of key frame detection in news video is shown in Fig.4.

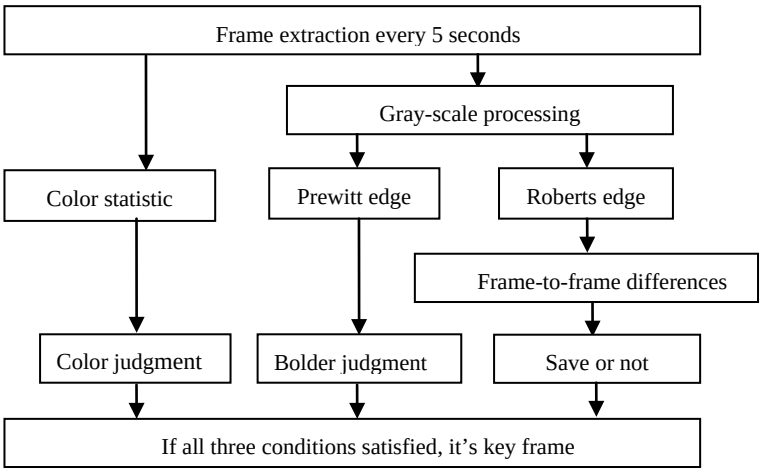

Fig. 4 The procedure of key frame detection

3.3 Locating of caption

After comparing several algorithms of caption region localization, we find that the algorithm based on edge detection is easy to locate captions. But when the caption background color is similar to the frame color, the detection accuracy is lowered; In addition, the algorithm based on Fuzzy C-Means clustering is hard to find the appropriate initial cluster centres.At the same time the effectiveness is influenced by the caption color information, so it's not suit for all kinds of news video. Considering the location of captions in one kind of news video is fixed, we propose an easy method for one specific kind of News video, which is getting the location by experiments, and it will fit easily for other news video by simply modifying the location.

3.4 caption segmentation

Caption segmentation is to divide the caption into two unique regions: texts and background. Among all the image segmentation algorithms, we find Otsu algorithm is appropriate for caption segmentation [9].

If the grayscale of one image is L, pixel gray value is［1,2,…,L］,the number of pixels whose gray value is i is n_i, so the number of pixels is

$$N = \sum_{i=0}^{L} n_i \tag{4}$$

Probability of pixels with a gray value of i is

$$P_i = n_i / n \tag{5}$$

Separate the image pixels with a gray threshold value of T into two categories: one class is the pixels with the grayscales of [0,….,T], denoted as D_0, the other class is the pixels with the grayscales of [T+1,…..,L], denoted asD_1.The probability of D_0 and D_1 are $P_0(T)$and $P_1(T)$, the grayscale average of D_0 and D_1 is $\mu_0(T)$ and $\mu_1(T)$.The variance of D_0 and D_1 are $\delta_0^{\,2}(T)$ and $\delta_1^{\,2}(T)$.The gray value of the whole image can be formulated as

$$\mu = \sum_{i-0}^{L} p_i = P_o(T)\mu_0(T) + P_1(T)\mu_1(T) \tag{6}$$

The squared distance between the two classes can be formulated as

$$\delta_b^{\,2}(T) = P_0(T)(\mu_0(T) - \mu)^2 + P_1(T)(\mu_1(T) - \mu)^2 \tag{7}$$

In order to improve processing speed and at the same time combine the results, the formulate can be simplified as

$$\delta_b^{\,2}(T) = \frac{(\mu_0(T) - \mu_1(T))^2}{P_0(T)P_1(T)} \tag{8}$$

The result is shown in Figure 5.

(a) Grayscale

(b) Binary image

Fig. 5 The result of Otsu algorithm

4. Results and Analysis

In order to test the effectiveness of the algorithm, five CCTV News programs are selected to be tested.

The recall (R) and precision (P), which are defined as follow, are used to evaluate performance of the proposed method.

$$P = \frac{R_A}{R_A + R_B} \tag{9}$$

$$R = \frac{R_A}{R_A + R_C} \tag{10}$$

Where R_A , R_B , R_C indicate the number of total key frames, the number of error detected frames, and the number of missed frames respectively. The result is shown in Table 1.

After caption segmentation, the binary captions are taken into OCR software for recognition. The result of recognition is shown in Table2.

As shown in Table 1, no frame of the detected video is missed in key frame detection, yet errors exist. The reason lies in cases where there is no caption in the frame, yet other region contains borders after edge detection. In caption text recognition, the average recognition rate is 87.3%. The error is always at the end of caption where the background color is similar to the frame.

Table 1 The result of key frame detection

Sequence number	Total frame number	Key frame number	Detected frame number	Missed frame number	Error frame number	Recall P	Precision R
1	13975	20	21	0	1	95%	100%
2	13825	21	24	0	3	87.5%	100%
3	14325	13	16	0	3	80%	100%
4	16600	18	19	0	1	94.7%	100%
5	17300	19	20	0	1	95%	100%

Table 2 The result of text extraction

Sequence number	Number of words	False detected words	Recognition rate
1	253	32	87.35%
2	296	30	89.86%
3	166	14	84.34%
4	275	35	87.27%
5	261	32	87.74%

5. Conclusion

Many algorithms of caption detection in news video have been proposed in recent years. But the difference of caption color, size and location between different news programs makes it difficult to obtain an efficient approach for all the news video. While the approach in this paper is fit for one news program with a set of parameters, it will fit easily for other news programs by simply modifying the parameters.

Acknowledgments: This work is supported by "863" national project, No. 2012AA01172

References

1. X. H. Huang, H. D. Ma "Automatic Detection and Localization of Natural Scene Text in Video" 2010 International Conference on Pattern Recognition .2010.786.
2. X. Zhao, K. H. Lin, and Y. Fu. "Text From Corners: A Novel Approach to Detect Text and Caption in Videos" IEEE transactions on image processing, vol. 20, No. 3, Mar. 2011.pp. 790-799.
3. M. Leon, V. Vilaplana, A. Gasull and F. Marques. "Region-based Caption Text Extraction, "Proc. IEEE Symp.2010 11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS). IEEE Press. Nov. 2010. pp 1-4.
4. Sharma, N.Shivakumara, P,Pal, U. ; Blumenstein, M,Tan, C.L. "A New Method for Arbitrarily-Oriented Text Detection in Video" 2012 10th IAPR International Workshop on Document Analysis Systems (DAS),10.1109/DAS.2012.6
5. B.Cai, D.R.Zhou, H.B.Hu "the study and implementation of caption detection and extraction in digital video" Journal of Computer-Aided Design&Computer Graphics 2003,15(7):898-903.
6. T.Sato,T.Kanade, E.K.Jughes, M.A.Smith and S.Satoh. "Video OCR：indexing digital news libraries by recognition of superimposed captions "ACM Multimedia Systems: Special Issue on Video Libraries. Vo1.7.N0.5, 1999, PP.385-395
7. M.Li, B.C.Li, D.W.Su. "Caption Detection and text content extraction algorithm “Video Engineering 1002-8692（2005）08-0147-03.
8. H.Su, H.X.Zhou, Z.H.Li.”The study of edge detection in image processing [J]” .Computer Development & Applications．2002,15(10):7-9.
9. L.N.Qi, B.Zhang, Z.K.Wang " The application of Otsu in image processing " Radio Engineering 2006.36（7）