Exploring Inter-Frame Correlation Analysis and Wavelet-Domain Modeling for Real-Time Caption Detection in Streaming Video

Jia Li*[a,b], Yonghong Tian [c], Wen Gao [a,c]
[a]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Science
[b]Graduate University of Chinese Academy of Sciences
[c]The Institute of Digital Media, Peking University

ABSTRACT

In recent years, the amount of streaming video has grown rapidly on the Web. Often, retrieving these streaming videos offers the challenge of indexing and analyzing the media in real time because the streams must be treated as effectively infinite in length, thus precluding offline processing. Generally speaking, captions are important semantic clues for video indexing and retrieval. However, existing caption detection methods often have difficulties to make real-time detection for streaming video, and few of them concern on the differentiation of captions from scene texts and scrolling texts. In general, these texts have different roles in streaming video retrieval. To overcome these difficulties, this paper proposes a novel approach which explores the inter-frame correlation analysis and wavelet-domain modeling for real-time caption detection in streaming video. In our approach, the inter-frame correlation information is used to distinguish caption texts from scene texts and scrolling texts. Moreover, wavelet-domain Generalized Gaussian Models (GGMs) are utilized to automatically remove non-text regions from each frame and only keep caption regions for further processing. Experiment results show that our approach is able to offer real-time caption detection with high recall and low false alarm rate, and also can effectively discern caption texts from the other texts even in low resolutions.

Keywords: Caption Detection, Inter-Frame Correlation, Generalized Gaussian Model, Streaming Video Analysis

1. INTRODUCTION

Recently, the amount of streaming video has grown rapidly on the Web. Different from video files in Web pages, streaming video are shown to viewers in the form of data-streams. To meet the increasing demand of finding a specific video stream from thousands of live broadcasts, the indexing and retrieval of streaming video becomes indispensable. However, existing Web video search engine cannot effectively retrieve such streaming videos for the reason that retrieving streaming videos offers the challenge of indexing and analyzing the media in real time, thus precluding offline processing [18]. To solve this problem, novel approaches are needed for real-time streaming video content analysis.

Generally, texts in video are important semantic clues for video indexing and retrieval. There are already a lot of researches on text detection in images and videos [1-10], [12-17], [20-21], [23-28]. Usually, these methods only concerned on the accurate localization of text regions and few of them stressed real-time processing. Moreover, these methods seldom concerned on how to separate caption texts from scene texts and scrolling texts. These texts usually play different roles in streaming video retrieval. For example, caption texts are usually about the current topic of the video clip, scene texts usually provide clues for natural scene understanding, while scrolling texts usually concern on other topics such as the latest news, commercials, supplementary information and so on. By distinguishing them effectively, streaming videos can be more accurately indexed from various aspects so that users can access a specific streaming video for different purposes. In addition, some methods such as [2], [9], [16] had explored the temporal information for verification and tracking of texts and thus could be used to check the types of text regions. However, these methods might spend extra time on the detection of scrolling texts and scene texts if only caption texts are needed in some applications such as video topic extraction.

To deal with these problems, we propose a novel approach for real-time caption detection in streaming video. For simplification, we only focus on the detection of caption texts in this paper. However, scene texts and scrolling texts can

*jli@jdl.ac.cn; phone: (86)010 6275 4541; www.jdl.ac.cn

be detected in a similar way. In our approach, we first explore the inter-frame correlation analysis to distinguish caption texts from scene texts and scrolling texts. After that, wavelet-domain Generalized Gaussian Models (GGMs) are utilized to automatically remove non-text regions and only retain caption regions for further processing. Finally, the inter-frame correlation coefficients (IFCCs) are used to provide robust and fast caption tracking and a Support Vector Machine (SVM) is adapted as the classifier to find caption regions in real-time.

This paper is organized as follows: Section 2 presents a simple survey of related works. Section 3 discusses the overall system architecture for real-time streaming video caption detection. Section 4 describes the proposed approach about inter-frame correlation analysis and wavelet domain modeling. Section 5 discusses the experiments and results. Finally, the paper is concluded in section 6.

2. RELATED WORK

There are many text detection algorithms and these algorithms can be generally classified into three categories according to the features used. In the first category, edge (gradient) is a preferred feature. In [1], Liu *et al.* use four Sobel edge detectors to obtain four edge maps and then 24 features are extracted in a sliding window. Based on these features, the window is classified into background and text candidates by using k-means algorithm. Finally, text areas are identified by the empirical rules analysis and refined through project profile analysis. In [16], Lyu *et al.* use Sobel detector to get enhanced edge map of an image, where the Sobel detectors consists of four directional gradient masks. Then by local thresholding and hysteresis edge recovery, text regions are accurately localized by several horizontal and vertical projections. Wang *et al.* [23] also use Sobel edges extracted from the image formed by multi-frame integration. Then a block based method is used to classify candidate text blocks as text or non-text by using the edge density in that block.

Besides spatial gradient detectors, in the second category, wavelet transform is also a frequently used tool to get high frequency points. Li *et al.* [9] implement a scale-space feature extractor that uses an artificial neural processor to detect text blocks. They also propose a text tracking scheme which uses a sum of squared difference based module to find the initial position and a contour based module to refine the position. Ye *et al.* [21] propose a coarse-to-fine strategy to detect texts in images and video frames. An image is first decomposed into several levels using wavelet transform. In coarse detection, wavelet energy features and density-based region growing method are used to get candidate text regions, which are then classified as texts or non-texts using a SVM classifier in the fine detection. This algorithm achieves a good performance but it does not utilize the temporal information. Due to the use of SVM classifier based on complex features, the algorithm is quite time-consuming. Wang *et al.* [28] propose a spatial-temporal wavelet transform for video text detection. Texture features are extracted from the combination of subbands of the spatial-temporal wavelet transform, and a Bayesian classifier takes charges of distinguishing text blocks from backgrounds. The algorithm can detect not only caption texts but also scene texts.

Moreover, in the third category, features extracted directly from the gray image are also used to find text regions. Kim *et al.* [14] propose a texture based method using a SVM classifier. They simply use the gray values of the points in the neighborhood of a pixel to train the SVM model and to classify that pixel as text or non-text pixel. Then the text regions are formed based on these text pixels using an adaptive mean shift (CAMSHIFT) algorithm. Although such features are easy to extract, it is insufficient for the classifier to discern text from complex background. Moreover, the pixel based algorithm is quite time-consuming and is not suitable for real-time text detection.

In conclusion, we can see that though a lot of text detection algorithms were proposed, none of them concern on the differentiation of caption texts, scene texts and scrolling texts. Furthermore, their computational complexities make them unsuitable for real-time text detection. As mentioned before, real-time detection performance is important for streaming video retrieval. Thus we still need to study real-time text detection algorithms with the ability of distinguishing text types.

3. SYSTEM ARCHITECTURE

The flowchart of our real-time caption detection system for streaming video is shown in figure 1. The system contains two main modules: temporal analysis module and spatial analysis module. Temporal analysis module is used to distinguish caption texts from scene texts and scrolling texts by using inter-frame correlation information. In spatial analysis module, global statistic information is used to automatically remove non-text regions. After that, captions will be detected from the regions left by using morphological operations, text tracking and a SVM classifier.

To utilize inter-frame correlations while making real-time caption detection, we use frame *i*-1 and *i*+1 when detecting captions in frame *i*. The abbreviation [*i*, *i*+1] is used to express the frame pair *i* and *i*+1. At the beginning, all the three frames are decomposed respectively into four wavelet subbands: LL, LH, HL and HH. Since text regions are usually full of strong edges and subband HH usually contains a lot of noises, only subbands LH and HL are used to compute inter-subband correlation coefficients (ISCCs). In this paper, WS is used to denote wavelet subband, $WS \in \{LH, HL\}$.

In our system, ISCC is ranged in [-1, 1]. For a point in WS , ISCC is used to represent the global stability of its neighboring edges in WS between consecutive frames. A larger ISCC means neighboring edges in WS are more stable between consecutive frames. To get a direct view, we scale all of the ISCCs to the range [0, 255] in this paper. Based on the ISCCs of subbands LH and HL between frame pairs [*i*-1, *i*] and [*i*, *i*+1], the inter-frame correlation coefficients (IFCCs) of frame pairs [*i*-1, *i*] and [*i*, *i*+1] are calculated to evaluate the global correlations of all points between consecutive frames. By combining these IFCCs, the Temporal Stability (TS) Map is finally obtained, where TS of a point in frame *i* denotes its max inter-frame correlation coefficients with the corresponding points in frame *i*-1 and frame *i*+1. The TS values are ranged in [-1, 1] and they are also scaled to [0, 255] to get a direct view in this paper.

To distinguish caption texts from scene texts and scrolling texts, TS may be a useful feature. Usually, scrolling texts move at a constant speed to a fixed direction, while scene texts will move due to the motion of cameras. That means points in scene text regions and scrolling text regions usually suffer smaller TS. Since caption texts are usually static between frames, TS values of caption regions will be relatively large. By applying a simple thresholding operation to the TS Map, points in scene text regions and scrolling text regions will be distinguished from points in caption text regions. After that, frame *i* is divided into two sub-images. For simplification, we focus on the detection of captions in the sub-image that contains only caption edges. However, scene texts and scrolling texts can be detected in the other sub-image using the same way. In figure 1, the sub-image that contains only caption edges is described with a binarized TS Map. The white points denote points in the TS Map of frame *i* whose TS values are larger than the pre-defined threshold.

To automatically remove background regions in the selected sub-image, two GGMs are established respectively for the subbands LH and HL of frame *i*. Based on the two models, two adaptive thresholds are automatically derived to remove non-text regions from the sub-image. After that, post processing methods are used to extract captions from the left regions in this sub-image. In the post processing step, morphological operations are first applied to segment the left regions into candidate caption lines. Then the tracking process will be carried out by exploring the IFCCs between frame pair [*i*-1, *i*]. After the tracking process, a SVM classifier is adapted to mark candidate caption lines left as ”*caption* ”or ”*non-text”*. The candidate caption lines marked as “*caption”* are mapped back to the original frame *i* so as to identify the location of the newly emerging captions in this frame.

To explicate the way to utilize temporal and spatial information in the caption detection, inter-frame correlation analysis and wavelet-domain modeling will be further explained in the following section.

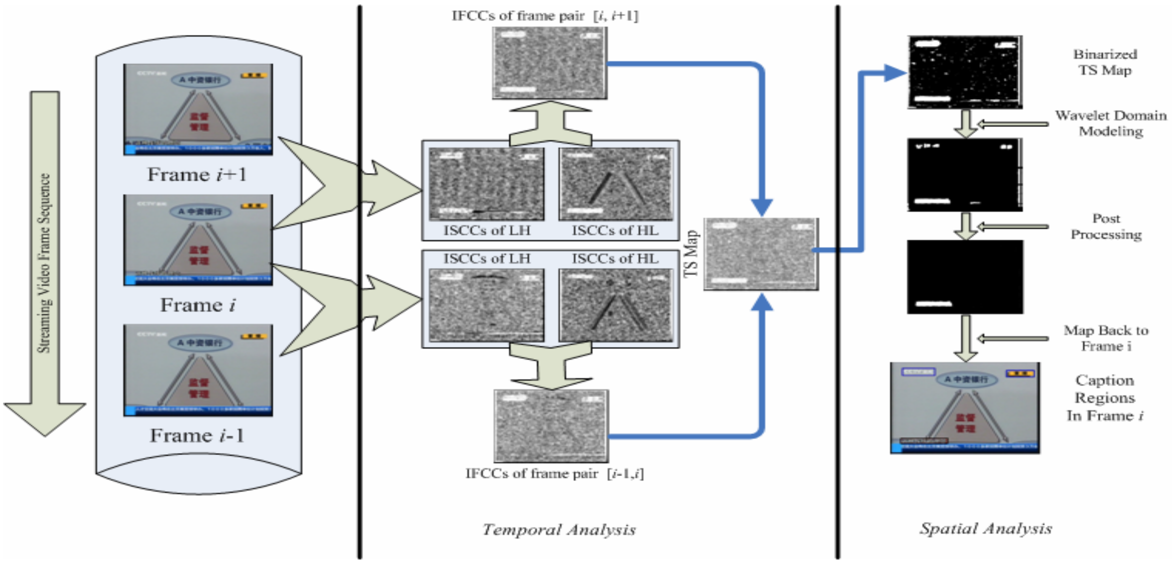


Fig. 1. Overall system architecture

4. THE ALGORITHM

To solve the problems for caption detection in streaming video, we propose a novel approach which utilizes temporal information before detecting captions in spatial domain. Firstly the inter-frame correlation analysis is explored to distinguish caption texts from scene texts and scrolling texts. Then wavelet subbands are modeled with GGMs and two adaptive thresholds are automatically derived to effectively remove non-text regions. Finally, morphological operations, text tracking and a SVM classifier are adopted to find captions in real-time. More details are discussed as follows.

4.1 Inter-Frame Correlation Analysis

To speed up caption detection, a reasonable sampling rate along temporal axis is required. Usually, caption texts are supposed to last at least 1 second and thus in this paper the sampling rate is fixed to 2 frames per second. This hypothesis has been validated by our experiments. An exception occurs in the case that the caption is very short and lasts only quite a short time. But these captions usually carry little information about the video content and thus can be ignored in the caption detection.

To discern caption texts from scene texts and scrolling texts in low resolutions, motion information could be useful. Scrolling texts are usually moving at a constant speed to a fixed direction, while scene texts will move due to the motion of cameras. Thus the temporal stability of edges may be a robust feature to discern them from caption texts whose edges are usually stable between frames. To effectively and efficiently evaluate the temporal stability in low-resolution streaming videos, we define the ISCC of point (x, y) in wavelet subband WS between frame pair [i-1, i] as follows:

$$\mathrm{ISCC}(x,y,i-1,i,WS)=\begin{cases}-1 & \bar{\sigma}_{i-1}(x,y,WS)\bar{\sigma}_{i}(x,y,WS)<\varepsilon,\\ MAX(MIN(1,\dfrac{\bar{\sigma}_{[i-1,\,i]}(x,y,WS)}{\bar{\sigma}_{i-1}(x,y,WS)\bar{\sigma}_{i}(x,y,WS)}),-1) & elsewise,\end{cases} \tag{1}$$

$$\bar{\sigma}^2_i(x,y,WS)=\frac{1}{(2M+1)^2}\sum_{a=x-M}^{x+M}\sum_{b=y-M}^{y+M}WS_i(a,b)^2\ , \tag{2}$$

$$\bar{\sigma}_{[i-1,\,i]}(x,y,WS)=\frac{1}{(2M+1)^2}\sum_{a=x-M}^{x+M}\sum_{b=y-M}^{y+M}WS_{i-1}(a,b)WS_i(a,b)\ , \tag{3}$$

Where $\bar{\sigma}^2_i(x,y,WS)$ and $\bar{\sigma}_{[i-1,\,i]}(x,y,WS)$ are the local variance and local covariance at point (x, y) in wavelet subband WS respectively. ε is a predefined small value to avoid the case of dividing by zero.

In formula (1), the predefined threshold ε is set to 0.1 so as to avoid the case of dividing by zero. In the computation, the nearby $(2M+1)\times(2M+1)$ coefficients are used to calculate the ISCC at point (x, y) in WS. It makes ISCC more stable since it is decided by not only the coefficient at point (x, y) but also coefficients around that point. In addition, the ISCC is restricted to the range [-1, 1]. In general, a larger ISCC at a point means that edges in its $(2M+1)\times(2M+1)$ neighboring region are more stable between frames. If $\mathrm{ISCC}\le 0$, the corresponding points in WS of frame i-1 and i are considered as completely irrelevant. In this paper, M is set to 3.

In order to quantify the global correlation of all points between frame pair [i-1, i], the IFCC is calculated as follows:

$$\mathrm{IFCC}(x,y,i-1,i)=\underset{WS\in\{LH,HL\}}{MAX}\{\ \mathrm{ISCC}(x,y,i-1,i,WS)\ \}\ , \tag{4}$$

To evaluate the max inter-frame correlation of all points in frame i, Temporal Stability (TS) of a point (x, y) in frame i is defined as follows:

$$\mathrm{TS}(x,y,i)=MAX\{\ \mathrm{IFCC}(x,y,i-1,i),\ \ \mathrm{IFCC}(x,y,i,i+1)\ \}\ , \tag{5}$$

For a point in frame i, its TS value is used to denote the max inter-frame correlation coefficients with the corresponding points in frame i-1 and frame i+1. By using TS, the stability of a point is quantized to the range [-1, 1]. After that, a predefined threshold $\mathrm{T_{TS}}$ is used as the criteria to divide frame i into two sub-images I_{H} and I_{L}, where I_{H} (or I_{L}) contains points whose TS values are larger (or smaller) than $\mathrm{T_{TS}}$ respectively. In the following processing, we focus on I_{H} which contains only caption edges while scene texts and scrolling texts can be detected in I_{L} using the same way. By experiment, $\mathrm{T_{TS}}$ is set to 0.6.

In the algorithm, the calculation of local variance and local covariance with a $(2M+1)\times(2M+1)$ template has ensured the robustness of ISCC. As shown in figure 2, the spatial sampling affects ISCC slightly. Thus it is possible to resize the image to a smaller version so as to get a faster detection speed while not remarkably degrading the analysis results.

In figure 3, it can be seen that changing background will not affect ISCC remarkably because edges of caption texts are usually stronger than background edges. Thus the inter-frame correlation coefficients of both caption and background points in the caption regions are mainly decided by the strong caption edges, not the edges of the background.

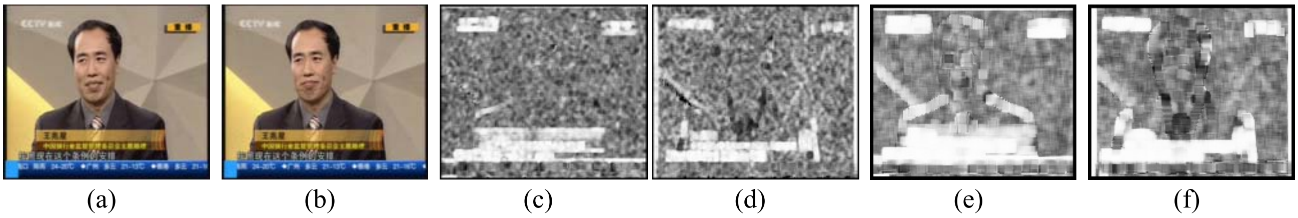


Fig. 2. ISCC Map under different resolutions. (a) and (b): original frame i-1 and i, 720*576;(c) and (d):ISCC Maps of LH and HL in resolution 720*576;(e) and (f): ISCC Maps of LH and HL in resolution 360*288

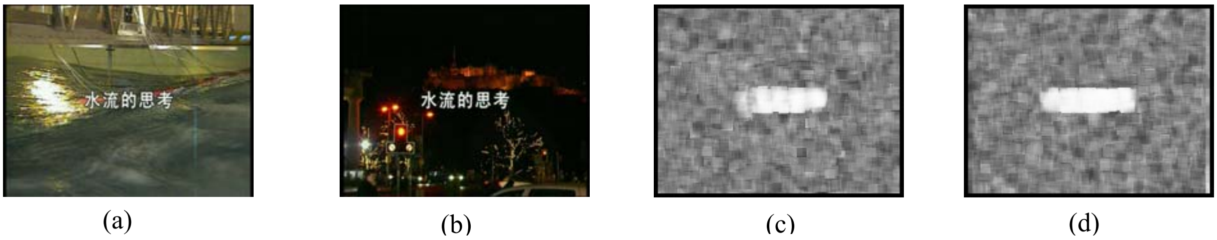


Fig. 3. Changing background affect ISCC slightly. (a) and (b):original frame *i*-1 and *i*;(c) and (d):ISCC of LH and HL.

The algorithm is also sensitive to slight motions. To further illustrate this point, a text region with simple background is selected out. This region displaces by *x* pixels to the right and forms a new text region. The mean IFCC of all points in the overlapped area of these two text regions are used to evaluate the sensitivity of the algorithm. As shown in figure 4, when *x* is not equal to zero the mean IFCC is very small. This means that the algorithm is quite sensitive to slight motion.

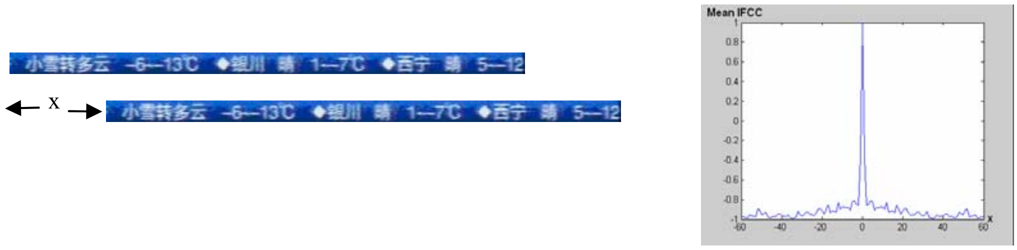


Fig. 4. IFCC is sensitive to slight motions

4.2 Wavelet-Domain Modeling

After inter-frame correlation analysis, only points in sub-image I_H that are relatively stable along temporal axis will be used for further analysis. For the shots with strong motion, only points in caption regions are left. But for shots with slight or no motion, especially those with a static background, a large number of non-text points are left in I_H. Thus in this paper, we use wavelet-domain Generalized Gaussian Models (GGMs) to automatically remove these non-text regions while retaining the caption regions.

As in [19], for natural images, the wavelet coefficient histograms in each subband are typically long-tailed, sharply peaked at zero and are commonly modeled by the Generalized Laplacian (also called Generalized Gaussian) Distributions. The Generalized Gaussian Distribution is given by the following formula:

$$f(x:\mu,\sigma^2,\gamma)=\mathrm{a}e^{-(\mathrm{b}|x-\mu|)^{\gamma}}\,, \qquad (6)$$

where μ,σ^2,γ are mean, variance, and shape parameter of the distributions respectively. $\Gamma(\bullet)$ is the gamma function given by $\Gamma(x)=\int_0^{\infty}t^{x-1}e^{-t}dt \quad x>0$ .The positive constants a and b are defined as: $\mathrm{a}=\frac{\mathrm{b}\gamma}{2\Gamma(1/\gamma)}$ and $\mathrm{b}=\frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}$,

Usually the GGM can be viewed as a simple way to express the wavelet coefficient histograms. In [11] a fast algorithm is proposed to find the optimal shape parameter γ .This paper experimentally selects an appropriate γ value from the interval [0.2, 2] with a step 0.01. Figure 5 summarizes the algorithm that is used to estimate the shape parameter γ .

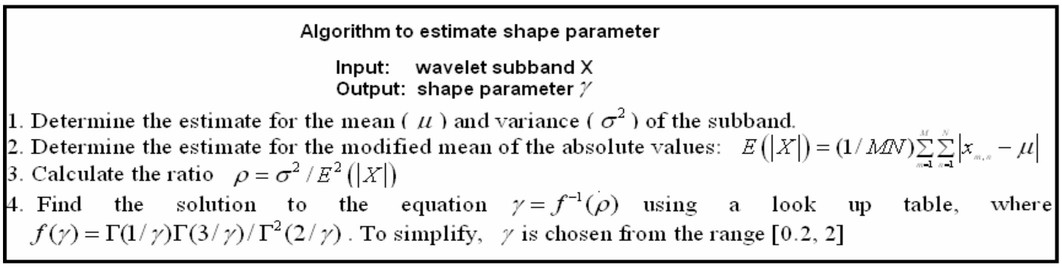
To remove the non-text points in I_{H}, both subbands LH and HL of frame *i* are modeled with GGMs respectively, and two adaptive thresholds are calculated by using μ,σ^2,γ of each model. Since subband coefficients can be considered as approximate zero-mean in most cases, the threshold for each subband is determined mainly by the variance and the shape parameter. As shown in figure 6, in the case of the same variance, a model with a smaller shape parameter means there are more weak edges in the subband, which consequently means a relatively ”clear” background. According to [16], on a clear background with a lot of weak edges, the threshold should be relatively low so as to let both low-contrast texts and high-contrast texts be detected. On a complex background, it should be relatively high so as to eliminate background and highlight texts. Thus the threshold should be proportional to the shape parameter and the variance. For wavelet subband WS the threshold is defined by:

$$threshold_{\mathrm{WS}}=\mathrm{c}\times\sqrt{\gamma_{WS}}\times\sigma_{WS} \qquad , \qquad (7)$$

The constant c is set according to experiments. Generally, a larger value of c will lead to a smaller recall and false alarm rate, and vice versa. In our experiments, c is set to 4.

In I_{H}, points whose absolute coefficients in each subband WS are smaller than $threshold_{WS}$ will be removed. After that, points with strong edges are left in sub-image I_{H} as ”Seed Point”.

In most cases, background points can be effectively removed by using $threshold_{\mathrm{WS}}$. But sometimes simple strong edges such as arrows and dividing lines are still left. To remove them, an algorithm is proposed to calculate the density of seed points. For a specific seed point, its density is defined as the number of seed points in its $(2M+1)\times(2M+1)$ neighborhood. Then seed points with densities less than $\mathrm{T_D}$ are removed. In order to effectively remove simple edges that usually contain two straight edge lines, $\mathrm{T_D}$ is set to $4M+2$. After that, only seed points whose densities are higher than $\mathrm{T_D}$ are kept in sub-image I_{H}. To form connected components, the points in the $(2M+1)\times(2M+1)$ neighborhood of these high-density seed points are also kept in sub-image I_{H}. These points will be further used in the post processing step.



Algorithm to estimate shape parameter

Input: wavelet subband X
Output: shape parameter γ

```
1. Determine the estimate for the mean ( $\mu$ ) and variance ( $\sigma^2$ ) of the subband.
2. Determine the estimate for the modified mean of the absolute values: $E(|X|)=(1/MN)\sum_{m=1}^{M}\sum_{n=1}^{N}|x_{m,n}-\mu|$
3. Calculate the ratio $\rho=\sigma^2/E^2(|X|)$
4. Find the solution to the equation $\gamma=f^{-1}(\rho)$ using a look up table, where
   $f(\gamma)=\Gamma(1/\gamma)\Gamma(3/\gamma)/\Gamma^2(2/\gamma)$ . To simplify, $\gamma$ is chosen from the range [0.2, 2]
```

Fig. 5. Fast Estimation of Shape Parameters

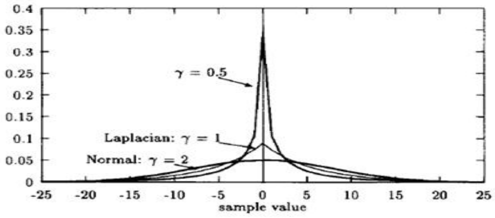


Fig. 6. Generalized Gaussian Distributions for different shape parameters but with the same variance

4.3 Post Processing

In the post processing, caption texts are detected from the regions left in sub-image I_H through the following procedure:

Step 1: Morphological operations. Adjacent regions are connected to form candidate text boxes. Then candidate text boxes are segmented into candidate text lines using the method proposed in [26].

Step 2: Inter-Frame tracking. The average IFCC between frame pair [*i*-1, i] are calculated for each candidate text line. Candidate text lines with high average IFCC are considered to be non-newly emerging captions and will be removed.

Step 3: SVM-based classification. For the candidate text lines left, a SVM classifier is used to classify them as ”*caption*“ and ”*non-text*”. To speed up the detection process, we use simple features such as wavelet moments and wavelet histogram proposed in [21].

After these steps, all caption texts in frame *i* are detected. In our algorithm, the following methods are used to ensure the real-time detection:

1) The algorithm is executed on two wavelet subbands, which makes the points to be processed reduce to 50%
2) When calculating ISCC, the variance and the covariance can be calculated using 2D-separable averaging filters.
3) The background can be effectively removed using global thresholds that are derived from GGMs, while the parameters of GGMs can be estimated using a simple algorithm. Compared with the local thresholding algorithm used in [16], global thresholds may greatly reduce the time cost.
4) In the post processing step, only simple features are extracted to feed the classifier. Compared with the complex features used in [21], simple features can effectively reduce the time cost, thus ensuring the real-time detection

By combining methods above, our algorithm can provide robust real-time caption detection while distinguishing caption texts from scene texts and scrolling texts.

5. EXPERIMENTAL RESULTS

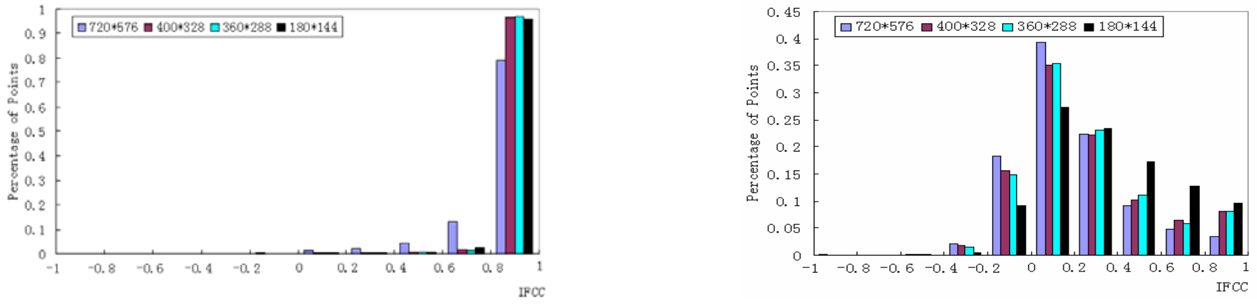
Four test sets are used in our experiments. The test set I contains 16 video clips with resolution 720×576, whose total length is about 6 hours 49 minutes. To evaluate the proposed algorithm on videos with different resolutions, all the 16 original video clips are degraded to videos with resolutions 400×328, 360×288 and 180×144, which are denoted as test set II, III and IV respectively. With the assumption that the duration of caption texts is at least 1 second, the sampling rate is fixed to 2 frames per second in our experiments. Thus, we obtain 49,177 frames with 89,639 caption regions from each test set. Daubichie4 wavelet transformation is employed in our algorithm for its good location performance. All the experiments are performed on a Pentium IV 3.2G CPU. Algorithms in [16] and [21] are realized for comparisons.

5.1 IFCCs of Text Regions

In the first experiment, we use all the four data sets to evaluate the robustness and sensitivity of inter-frame correlation coefficients. Specifically, IFCCs between adjacent frames, which contain the same texts in corresponding caption regions, are used to evaluate the robustness; while IFCCs between adjacent frames, which contain different caption texts, are used to evaluate the sensitivity. For instance, Let R_{i-1}, R_i, R_{i+1}… R_{i+N}, R_{i+N+1} denote the regions that locate at the same position in frame *i*-1, *i*, *i*+1 …*i*+N, *i*+N+1 respective. Among the N+3 regions, Only R_i, R_{i+1}… R_{i+N} contain the same caption texts. As shown in figure 7(a), the histogram of the IFCCs between regions [R_i, R_{i+1}]$_{\ldots\ldots}$ [R_{i+N-1},R_{i+N}] is calculated to evaluate the robustness. The histogram of IFCCs between regions [R_{i-1}, R_i] and [$R_{i+N,}$ R_{i+N+1}] is computed to evaluate the sensitivity and the result is shown in figure 7(b).

Since captions with same texts are on different backgrounds and their IFCCs in figure 7(a) are pretty high, we can conclude that IFCCs are robust to various backgrounds. Meanwhile, in figure 7(b), region pairs with different text contents are featured with small IFCCs, which demonstrate the sensitivity of IFCCs to caption changes. Thus IFCCs can effectively reflect the correlations between regions or frames.

Note that in figure 7(a), IFCCs in resolution 720×576 are smaller than those in low resolutions for the reason that the template with size $(2M+1) \times (2M+1)$ used in the calculation of IFCCs is relatively small in high resolution. In addition, the percentage of great IFCCs in resolution 180×144 is higher than that in other three resolutions due to the smooth or lost of caption edges and the emergence of unexpected background edges when degrading videos to low resolutions.



(a) (b)

Fig. 7. (a) IFCC histogram to evaluate robustness;(b) IFCC histogram to evaluate the sensitivity

5.2 Detection Speed

In this experiment, we use 7 frame sequences from data set II, which have the same resolution 400×328 with those used in [21], to evaluate the detection speed. To get fair comparisons, all the frames contain no scrolling texts and only few scene texts. Table 1 shows the detection speed of different algorithms, from which we can tell that our proposed algorithm significantly outperforms the methods in [16] and [21].

Nevertheless, in real-time video processing, high detection speed cannot necessarily guarantee accurate results. In some cases, complicated frames could take high computational cost and therefore leads to skipping of the following frames. Motivated by this fact, we introduce nonstationarity to evaluate the variance of detection speed. And it is formally defined as follows:

$$\text{Nonstationarity} = \frac{\sqrt{\sum_{i=1}^{7}\sum_{j=1}^{N_i}[(S_{ij} - S_{ALL})^2 / \sum_{i=1}^{7} N_i]}}{S_{ALL}}, \tag{8}$$

Where S_{ALL} denotes the average detection speed on all the frame sequences selected, N_i denotes the total number of images in the frame sequence of the i-th video, and S_{ij} denotes the detection speed on the j-th frame in the i-th video. Intuitively, detection with lower nonstationarity is more tolerant to the speed changes. The nonstationarity comparison of different detection algorithms are also shown in table 1.

From table 1 we can see that our algorithm achieves a smaller nonstationarity than other algorithms, owing to the effectiveness of background removal in our method. By exploiting the distribution of wavelet coefficients, our algorithm can automatically analyze the frames and effectively remove various backgrounds, which ensure a low nonstationarity.

Table 1. Detection Speed and Speed Nonstationarity

Algorithm	**Our**	**[16]**	**[21]**
Speed(Frames/s)	**9.09**	4.46	1.18
Nonstationarity (%)	**5.13**	11.54	12.69

5.3 Removal of Scene texts and Scrolling texts

In this experiment, we select 9 news video with different resolutions from the four data sets that are rich of scene texts and scrolling texts. Experimental results show that all scrolling texts, no matter with clear background or complex background, can be removed since they are always moving at a constant speed and are totally different from the static caption texts. Nonetheless, some scene texts in strictly static shots tend to have the same properties with caption texts and therefore are failed to be removed. Figure 8 gives some examples of correctly removed scene texts and scrolling texts. Figure 9 shows some failure examples. Generally, the scene texts in moving shots and scrolling texts can be easily removed. This can greatly reduce the time cost on caption detection. While for a static shot, the scene texts are usually featured with the same proprieties of the caption texts so that it is quite difficult to distinguish them.

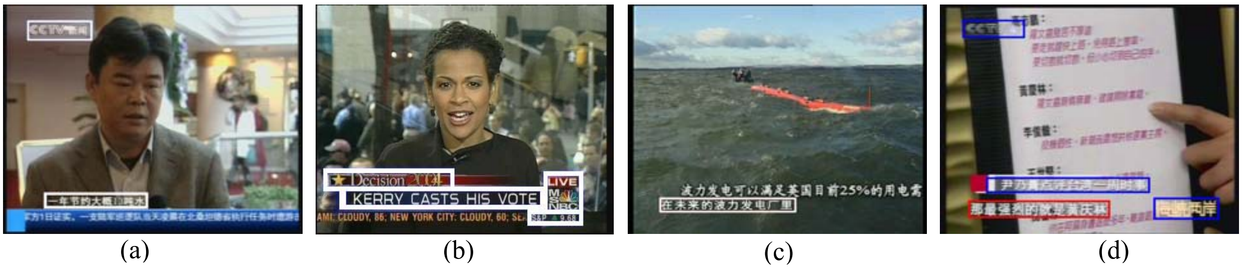


Fig. 8. Examples of successfully removed scene texts and scrolling texts.(a). Small scrolling texts on the bottom with simple background. (b). small scrolling texts on the bottom with transparent background. (c). Big scrolling texts above the caption with transparent background. (d). scene texts

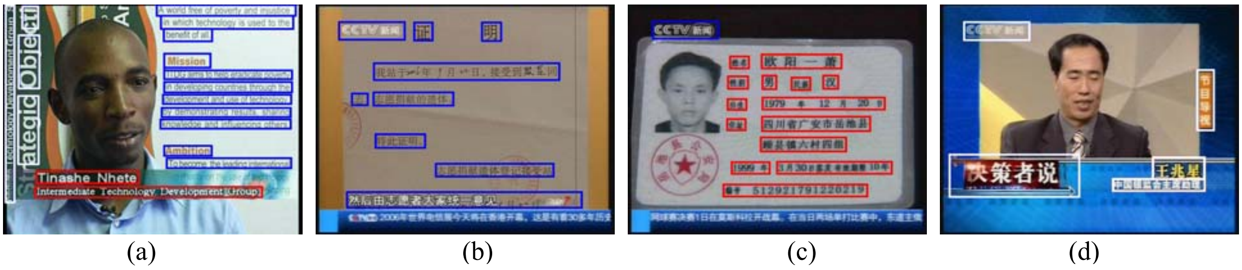


Fig. 9. Examples of failed removed scene texts.

5.4 Results of Caption Text Detection

For fair comparison, we select 7 videos from test set II containing 22,409 frames and 25,917 caption regions. Therefore, the selected frames have the same resolution with those in [21]. In order to focus on caption text detection, all the selected videos contain no scrolling texts and only few scene texts. Let S_G and S_D denote the area of ground-truth text regions and detected text regions respectively. Accordingly, the areas of their intersection are defined as S_O. The recall and false alarm rate are then defined as follows:

$$\text{Recall} = S_O / S_G \,, \tag{9}$$

$$\text{False alarm rate} = 1 - S_O / S_D \,, \tag{10}$$

As computational cost is the key issue in streaming video analysis and usually propotional to the areas of detected text regions in the following text processing steps (e.g. text segmentation and recognition), we adopt areas to calculate the recall and false alarm rate instead of the number of regions used in [15], [16], [21]. For instance, in figure 10, four regions are detected as caption text regions. Among them, three small regions are correct and a large region is falsely detected. In the view of computional cost, such detection results seems not quite satisfactory. Using the number of regions will lead to a false alarm rate of 25.0%, which cannot effectively reflect the computational cost wasted in the following step. By using the area as the critera in formula (10), the false alarm rate reaches 79.5%. This number can effectively reflect the extra computational cost wasted in the following text processing procedures. Thus using the area as the critera in real-time streaming video caption detection will surely make sense.

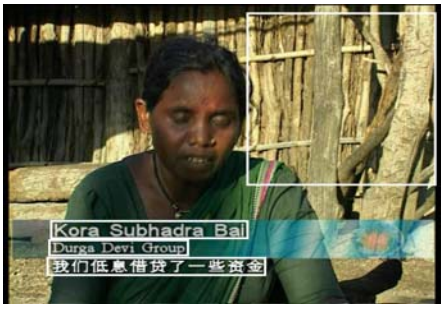


Fig 10. Illustration for the reason of the proposed evaluation metrics in caption detection

Based on the proposed evaluation strategy, two experiments are performed to compare our algorithm with the methods in [16] and [21]. In the first experiment, the 7 selected frame sequences are treated as off-line image sequences. Consequently, the detection is performed without time restriction. In the second experiment, the 7 selected frame sequences are streamed into the detection systems and the algorithms are required to generate real-time results. In addition, in the second experiment the temporal coverage, which equals to the ratio of processed frames to the total frames, is used to evaluate the performance under real-time restriction. To get a high temporal coverage, the algorithm must detect captions faster than 2 frames per second. The algorithm also has to provide a steady detection speed. Generally speaking, a steady detection speed for streaming video might avoid spending too much time on a single frame and consequently would not lead to skipping of the following several frames. The results of the two experiments are listed in table 2.

Through the results of the two experiments, our algorithm can reach the highest recall and the smallest false alarm rate. The main reason is that non-text regions can be effectively removed by the wavelet-domain modeling, and in the inter-frame correlation analysis some moving text-like textures are also removed. Compared with our method, the background-complexity adaptive local thresholding algorithm in [16] relies heavily on the predefined parameters. Although their algorithm can effectively highlight the text regions by enhance the text edges, they still require a robust classifier to distinguish texts from text-like textures. Generally, a local threshold makes it hard to ”suppress” the complex background edges by using strong text edges. In addition, our method also outperforms the algorithm in [21]. A possible reason is that the algorithm in [21] only selects a fixed percentage of pixels as candidate text edges in each frame, consequently leading to additional false alarms in no-caption frames and missing detections in rich-caption frames.

With a high and steady detection speed, our algorithm and algorithm [16] can give a 100% temporal coverage. This means every frame in the video frame sequences can be analyzed in real-time. It reveals that in real-time detection, the features used for distinguishing texts and non-texts cannot be too complicated since they may result in high computational cost. Consequently, it explains the reason why algorithm in [21] suffers a low temporal coverage, namely, the significant decrease in recall.

Table 2. Performance comparison in experiment 1 and experiment 2

	Experiment 1		**Experiment 2**		
Algorithm	**Recall (%)**	**False Alarm Rate (%)**	**Recall (%)**	**False Alarm Rate (%)**	**Temporal Coverage (%)**
Our	**90.66**	**28.98**	**90.66**	**28.98**	**100**
Algorithm [16]	82.11	38.17	82.11	38.17	**100**
Algorithm [21]	88.68	37.49	55.99	36.93	65.02

6. CONCLUSIONS AND FUTURE WORK

In this paper, we explore the inter-frame correlation analysis and wavelet domain modeling for real-time caption detection in streaming video. By using the inter-frame correlation analysis, the scene texts and scrolling texts can be distinguished from caption texts. After that, the wavelet subbands of current frame are modeled with two GGMs, from which two adaptive thresholds are derived to automatically remove backgrounds. This approach also performs well to remove non-text regions in frames that contain no caption texts. The inter-frame correlation coefficients can be calculated using a 2D-separable average filter and the parameters of the two GGMs can also be estimated using a fast algorithm. Moreover, there are few predefined parameters in the algorithm, which makes the caption detection algorithm fast and robust for various kinds of streaming video programs.

In future, we will explore better SVM features which can be easily extracted to decrease the false alarm rate. After that, the streaming video retrieval system will be further explored to index and retrieve live-streaming videos in real-time.

7. ACKNOWLEDGEMENT

This work is supported by grants from Chinese NSF under contract No. 60605020, National Hi-Tech R&D Program (863) of China under contract No. 2006AA01Z320 and 2006AA010105, and National Key Technology R&D Program under contract No. 2006BAH02A10.

REFERENCES

1. Chunmei Liu, Chunheng Wang, and Ruwei Dai. “Text detection in images based on unsupervised classification of edge-based features”. International Conference on Document Analysis and Recognition. 29 Aug.-1 Sept. 2005 Page(s):610 - 614 Vol. 2

2. Congjie Mi, Yuan Xu, Hong Lu, and Xiangyang Xue. “A Novel Video Text Extraction Approach Based on Multiple Frames”. International Conference on Information, Communications and Signal Processing, 06-09 Dec. 2005 Page(s):678 - 682

3. D. Chen, J.M. Odobez, and H. Bourlard. “Text detection and recognition in images and video frames”. Pattern Recognition, 37(3):595–608, 2004.

4. Dongqing Zhang, Rajendran, R.K., Shih-Fu Chang. “General and domain-specific techniques for detecting and recognizing superimposed text in video”. International Conference on Image Processing, 22-25 Sept. 2002. Volume 1, Page(s):I-593 - I-596

5. Dongqing Zhang, and Shih-Fu Chang. “Learning to Detect Scene Text Using a Higher-Order MRF with Belief Propagation”. Conference on Computer Vision and Pattern Recognition Workshop, 27-02 June 2004 Page(s):101 – 101

6. Ezaki, N., Bulacu, M., and Schomaker, L. “Text detection from natural scene images: towards a system for visually impaired persons”. International Conference on Pattern Recognition, Volume 2, 23-26 Aug. 2004 Page(s):683 - 686 Vol.2 .

7. F. Chang, G-C. Chen, C-C. Lin, and W-H. Lin. “Caption analysis and recognition for building video indexing systems”. ACM Multimedia Systems Journal, volume 10, number 4, pages 344-355, 2005.

8. Hongxing Sun, Nannan Zhao, and Xinhe Xu. “Extraction of Text under Complex Background Using Wavelet Transform and Support Vector Machine”. IEEE International Conference on Mechatronics and Automation,June 2006 Page(s):1493 – 1497

9. Huiping Li, Doermann David, and Omid Kia. “Automatic text detection and tracking in digital video. IEEE Transactions on Image Processing”, Volume 9, Issue 1, Jan. 2000 Page(s):147 – 156

10. Tran, H., Lux, A., Nguyen, T.H. L., and Boucher, A. “A novel approach for text detection in images using structural features”. International Conference on Advances in Pattern Recognition, Bath, 22-25 august 2005, vol. 3686,Page(s): 627-635

11. Kamran, S., and Alberto, L.G. “Estimation of shape parameter for generalized Gaussian distribution in subband decompositions of video”. IEEE Trans on circuits and systems for video technology. Volume 5, Issue 1, Feb. 1995 Page(s):52 – 56

12. Keechul Jung, Kwang In Kim, and Anil K Jain. “Text information extraction in images and video:a survey”. Pattern Recognition,2004,37(5):977-997.

13. Kim, K.C., Byun, H.R., Song, Y.J., Choi, Y.W., Chi, S.Y., Kim, K.K., and Chung, Y.K. “Scene text extraction in natural scene images using hierarchical feature combining and verification”. International Conference on Pattern Recognition, Volume 2, 23-26 Aug. 2004 Page(s):679 - 682 Vol.2

14. Kwang In Kim, Keechul Jung, and Jin Hyung Kim. „Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm”. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 25, Issue 12, Dec. 2003 Page(s):1631 – 1639

15. Lienhart, R., and Wernicke, A. “Localizing and segmenting text in images and videos”. IEEE Transactions on Circuits and Systems for Video Technology, Volume 12, Issue 4, April 2002 Page(s):256 – 268

16. Lyu, M.R., Jiqiang Song, and Min Cai. “A comprehensive method for multilingual video text detection, localization, and extraction”. IEEE Trans on circuits and systems for video technology, Volume 15, Issue 2, Feb. 2005 Page(s):243 – 255

17. Min Cai, Jiqiang Song, and Lyu, M.R. A new approach for video text detection. International Conference on Image Processing. Volume 1, 22-25 Sept. 2002 Page(s):I-117 - I-120 vol.1

18. Pieper, J., Srinivasan, S., and Dom, B.”Streaming-media knowledge discovery”. Computer,Volume 34, Issue 9, Sept. 2001 Page(s):68 – 74

19. Pizurica, A., Zlokolica, V., and Philips, W. Combined wavelet domain and temporal video denoising. Proceedings. IEEE Conference on Advanced Video and Signal Based Surveillance. 21-22 July 2003 Page(s):334 – 341

20. Qifeng Liu, Cheolkon Jung, and Youngsu Moon. “Text segmentation based on stroke filter”. Proceedings of the 14th annual ACM international conference on Multimedia Pages: 129 – 132

21. Qixiang Ye, Qingming Huang, Wen Gao, and Debin Zhao. “Fast and robust text detection in images and video frames”. Image and Vision Computing. Vol.23, No.6, pp565-576,Mar.2005

22. S. M. Mahbubur Rahman, M. Omair Ahmad, and M. N. S. Swamy. “Video Denoising Based on Inter-frame Statistical Modeling of Wavelet Coefficients”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 17, Issue 2, Feb. 2007 Page(s):187 – 198

23. Rongrong Wang, Wanjun Jin, and Lide Wu.“A novel video caption detection approach using multi-frame integration”,International Conference on Pattern Recognition, 2004,Volume 1, 23-26 Aug. 2004 Page(s):449 - 452 Vol.1

24. Silapachote, P., Weinman, J., Hanson, A., Mattar, M.A., and Weiss, R. “Automatic Sign Detection and Recognition in Natural Scenes”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 3, 20-26 June 2005 Page(s):27 – 27

25. Tsung-Han Tsai, Yung-Chien Chen, and Chih-Lun Fang .”A Comprehensive Motion Videotext Detection Localization and Extraction Method”, International Conference on Communications, Circuits and Systems Proceedings, 2006 Volume: 1, On page(s): 515-519

26. Xian-Sheng Hua, Liu Wenyin, and Hong-Jiang Zhang. “An automatic performance evaluation protocol for video text detection algorithms”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 14, Issue 4, April 2004 Page(s):498 – 507

27. Xilin Chen, Jie Yang, Jing Zhang, and Waibel, A. “Automatic detection and recognition of signs from natural scenes”, IEEE Transactions on Image Processing, Volume 13, Issue 1, Jan. 2004 Page(s):87 – 99

28. Yuan-Kai Wang, and Jian-Ming Chen. Detecting Video Texts Using Spatial-Temporal Wavelet Transform. International Conference on Pattern Recognition. Volume 4, 20-24 Aug. 2006 Page(s):754 - 757