A Spatial-Temporal Approach for Video Caption Detection and Recognition

Xiaoou Tang, *Senior Member, IEEE*, Xinbo Gao, *Member, IEEE*, Jianzhuang Liu, and Hongjiang Zhang, *Senior Member, IEEE*

***Abstract*—We present a video caption detection and recognition system based on a fuzzy-clustering neural network (FCNN) classifier. Using a novel caption-transition detection scheme we locate both spatial and temporal positions of video captions with high precision and efficiency. Then employing several new character segmentation and binarization techniques, we improve the Chinese video-caption recognition accuracy from 13% to 86% on a set of news video captions. As the first attempt on Chinese video-caption recognition, our experiment results are very encouraging.**

***Index Terms*—Chinese caption detection, fuzzy clustering neural networks (FCNNs), video indexing, video OCR, video shot segmentation.**

I. Introduction

THE ongoing proliferation of digital image and video libraries leads to an increasing demand for systems that can automatically query and search large video databases. To construct such systems, both low-level features such as object shape, region intensity, color, texture, motion descriptors, audio measurements, and high-level techniques such as human face detection, speaker identification, and character recognition have been studied for indexing and retrieving image and video information in recent years [3], [4], [10], [11], [13], [19], [21], [24], [27]–[29], [32], [36]. Among these techniques, video caption based methods have attracted particular attention due to the rich content information contained in caption text [1], [2], [6], [9], [11]–[13], [15], [16], [19], [20], [27], [33], [36]. Caption text routinely provides such valuable indexing information as scene locations, speaker names, program introductions, sports scores, special announcements, dates and time. Compared to other video features, information in caption text is highly compact and structured, thus is more suitable for efficient video indexing.

However, extracting captions embedded in video frames is not a trivial task. In comparison to OCR for document images, caption extraction and recognition in videos presents several new challenges [27]. First, captions in videos are often embedded in complex backgrounds, making it more difficult to be detected. Second, characters in captions tend to have a very low resolution since they are usually made small to avoid occluding scene objects in a video frame. Therefore, the quality of characters in a video frame is not suitable for direct processing in conventional OCR systems. In addition, popular lossy compression methods such as MPEG often lower the image quality even further.

In order to overcome these difficulties, new text detection and extraction methods have been developed recently. They are generally classified into three categories—connected component methods, texture classification methods, and edge detection methods. The connected component methods are based on the assumption that text is represented with a uniform color. After color quantization, the connected components of monotonous color that obey certain size, shape, and spatial alignment constraints are extracted as text [11], [13], [17], [20], [30], [39]. The methods are efficient when the background mainly contains uniform regions. But their efficiency tends to deteriorate when the background is cluttered.

The texture classification methods treat the text region as a special type of texture. Texture features extracted through multichannel processing [19], [36], spatial variance computing [39], and neural networks [12] have been used to detect text regions. In general, the texture-based methods are more robust than the connected component-based methods in dealing with a complex background. However, the texture-based methods find difficulties when the background contains textures that display similar periodic structures as the text region. In addition, given the large amount of data in digital videos, many sophisticated texture analysis methods cannot be used because of the computational complexity.

Recently, the edge detection methods have been increasingly used for caption extraction. Since characters are composed of line segments, text regions contain rich edge information [1], [2], [8], [16], [27], [33]. Utilizing the high-frequency information in text edges, the methods are effective in segmenting text from surrounding background after the text bounding box is already detected. For detection of the text region in a large video frame, they are less reliable since other objects and structures in the scene may also contain strong edges.

Many of the existing works deal with text detection in static images [8], [14], [17], [24], [26], [30], [35], [36], [39], [40]. Even though some address text detection in video frames [1], [5], [11]–[13], [16], [20], [34], they usually treat each video frame as an independent image. When temporal information are utilized, they are used only for text enhancement through multiframe averaging [18] or time-based minimum pixel search [15],

Manuscript received April 24, 2001; revised November 10, 2001. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region under Project CUHK4378/99E.

X. Tang and J. Liu are with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (e-mail: xtang@ie.cuhk.edu.hk).

X. Gao was with The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. He is now with the School of Electronic Engineering, Xidian University, Xi'an 710071, China.

H. Zhang is with Microsoft Research Asia, Beijing 100080, China.

Publisher Item Identifier S 1045-9227(02)04416-8.

[20], [27], [28]. These approaches require text detection and localization for every frame of a video, and careful caption blocks tracing and matching are needed between each frame pair for multiframe enhancement and removal of duplicate captions in different frames. Given the large number of video frames, the computation requirement can be very high. In this paper, we propose a new caption (dis)appearance detection method based on shot boundary detection and feature extraction from frame differences. By precisely locating the critical frame where each caption block appears or disappears, we can focus our caption detection operation solely on the critical frame difference. More salient text features can be extracted from the frame difference than from the original frame.

We also observe that many of the existing works require supervised training [12], [19], [33]. This not only puts a great deal of burden on the users but also renders the system much less flexible in dealing with different databases, since new training data have to be collected and new parameters have to be computed for each new database. To develop an unsupervised method, we use the self-organizing fuzzy clustering neural network (FCNN) as the classifier. Since the features we extract through caption (dis)appearance detection are more distinct than traditional caption features, the self-organizing classifier is much easier to converge.

Finally, the majority of the existing researches on caption extraction have been focusing on English captions, with a few exceptions addressing captions in Korean and Japanese [12], [16]. We have found few existing researches dealing with captions in Chinese. Since Chinese characters are more complex in structure and have far more categories (more than 5000) than English letters, recognition of low-resolution Chinese characters in a video frame is potentially more difficult. In this paper, we study the extraction and recognition of video captions in Chinese news videos. We choose news videos for two reasons: the importance of automatic indexing of the huge amount of video news programs covering daily events all over the world; and the large concentration of caption information presented in news programs.

We consider the special properties of Chinese characters in each processing step. For caption detection, we propose features that reflect the four types of primitive strokes and dense stroke patterns of Chinese characters. Then, for character extraction and segmentation, we take advantage of the rectangular shape and aspect ratio of Chinese characters to separate individual characters. Finally, we try to recognize the extracted characters by using a commercial Chinese OCR package. Given the large number of existing Chinese OCR algorithms and commercial Chinese OCR software packages, we do not need to develop any new OCR algorithms. However, we do consider the recognition step essential to determine whether the extracted captions are recognizable, and thus to verify the efficacy of the entire system.

The remainder of this paper is organized as follows. In Section II, an FCNN model is first described as an important processing tool in our system. Then we present the caption detection algorithm based on FCNN, including video shot segmentation, caption transition detection, and caption location. Chinese character enhancement and recognition are described in Section III. Section IV presents the experimental results and analysis. Finally, we conclude this paper in Section V.

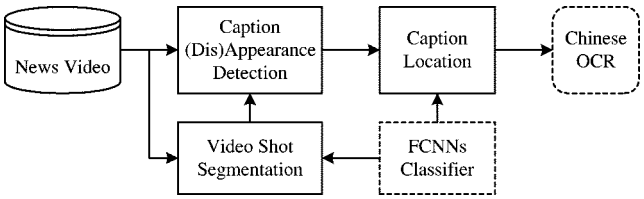


Fig. 1. The block diagram of the caption detection scheme.

II. Caption Detection

In order to extract captions from video frames, we need first find out the frames with captions then locate the caption regions. Normally, in a news video program, not all frames have captions and in the frames that do have captions, the same caption text usually appears over a multiple number of frames for stable viewing. In many of the previous researches, caption detection is usually carried out for every frame of the video sequence, then matching of caption regions between neighbor frames is conducted to eliminate duplicate captions. This process is time consuming because even a short segment of video has a large number of frames. Since captions are mostly superimposed on video frames manually, they appear and disappear from a video sequence more rapidly than regular scene content in a shot. Based on this observation, we propose a new scheme for detecting the caption (dis)appearance frame pair and locating the caption regions in the frame difference.

The scheme consists of three components as illustrated in Fig. 1. First, the video sequence is segmented into camera shots using a fuzzy clustering neural-network classifier. Then within a camera shot, we can assume that the scene changes gradually. By analyzing the differences between adjacent frames, we find that the (dis)appearance of caption text introduces a change much greater than most of the gradual-changing scene content. For the critical frame pair where caption text (dis)appears, the caption text regions are much more pronounced in the inter-frame differences than in the original video frame, thus can be more easily located. By focusing only on caption transition frames, we can avoid conducting caption detection for every video frame. Finally, the same FCNN classifier for shot boundary detection can be used for locating caption regions in the transition frame difference.

In the rest of this section, we first describe the FCNN classifier used in the caption detection scheme, then we discuss the detail approaches in each step.

A. FCNN Classifier

We choose a self-organizing neural network as the classifier because the network can adapt to different datasets fairly easily with little human intervention, which is critical toward developing a fully automated system [22], [25], [37]. In particular, since there are no sharp boundaries between different feature classes in our system, we choose a fuzzy clustering network as the classifier.

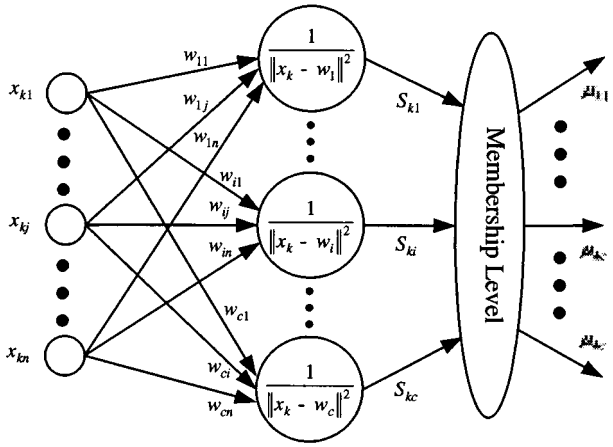


Fig. 2. FCNN model.

In the FCNN, one tries to classify the samples into clusters according to the natural structure hidden in the unlabeled n-dimensional data. Let $X = \{x_1, x_2, \ldots, x_N\} \subset \Re^n$ denote the samples at hand, and let c denote the number of nodes (clusters in X) in the competitive layer, a basic FCNN model using clustering competitive network is illustrated in Fig. 2. The input layer of the FCNN is connected directly to the competitive layer. The connecting weights from the inputs to a node, $w_i = (w_{i1}, w_{i2}, \ldots, w_{in})$, $1 \le i \le c$, represent the prototype of that fuzzy cluster. The output of the node, S_i, $1 \le i \le c$, represents the similarity of the input pattern to the prototype. For an input vector x_k, the S_{ki} is defined as

$$S_{ki} = \frac{1}{\|x_k - w_i\|^2} = \frac{1}{\sum_{j=1}^{n} (x_{kj} - w_{ij})^2}. \tag{1}$$

To avoid the singular case, a very small positive number η is often added to the denominator of the above formula. The membership value that describes the degree of each input pattern belongs to a fuzzy cluster is calculated in the final membership level by

$$\mu_{ki} = \frac{S_{ki}}{\sum_{l=1}^{c} S_{kl}} = \frac{1}{\sum_{l=1}^{c} \left(\frac{\|x_k - w_i\|}{\|x_k - w_l\|} \right)^2} \tag{2}$$

where $\mu_{ki} \in [0,1]$ and $\sum_{i=1}^{c} \mu_{ki} = 1$.

To train the FCNN, an energy function Q is defined as half of the mean-square-error summation between the input patterns and prototype vectors

$$Q = \frac{1}{2} \sum_{k=1}^{N} Q_k = \frac{1}{2} \sum_{k=1}^{N} \sum_{i=1}^{c} (\mu_{ki} \cdot \|x_k - w_i\|)^2 . \tag{3}$$

The energy function reaches minimal when each prototype weight vector is closest to the center of the corresponding sample cluster. To minimize the energy function, the gradient descent method is used to adjust the weight vectors. For a given input training sample, the gradient of the energy function with respect to the weight vectors is

$$\begin{aligned} \frac{\partial Q_k}{\partial w_i} &= \frac{\partial}{\partial w_i} \left\{ \frac{1}{2} \sum_{l=1}^{c} (\mu_{kl} \cdot \|x_k - w_l\|)^2 \right\} \\ &= \frac{1}{2} \frac{\partial}{\partial w_i} \left\{ (\mu_{ki} \cdot \|x_k - w_i\|)^2 \right\} \\ &= -(\mu_{ki})^2 \cdot (x_k - w_i). \end{aligned} \tag{4}$$

Since the negative gradient of the energy function points to the direction which will most quickly reduce the energy function, based on (4), the update rule of weight vectors can be formulated as

$$w_i^{(t)} = w_i^{(t-1)} + \alpha_t \cdot \left(\mu_{ki}^{(t-1)} \right)^2 \cdot \left(x_k - w_i^{(t-1)} \right) \tag{5}$$

where α_t is the correction factor. To avoid possible oscillations of the solution, the amount of correction should be reduced as iteration proceeds. Here we adopt Pal's scheme and take α_t as $\alpha_0(1 - t/T)$, where T is the maximum number of iterations that the learning process is allowed to execute and α_0 is the initial value of the learning parameter [25]. In each iteration, all training samples are passed through the network once. The training stops when the weight update for an iteration is smaller than a preset threshold.

B. Video Shot Segmentation

In the first step of the caption detection scheme, we use the FCNN to segment the video sequence into camera shots, which are the basic video units representing a continuous action in both time and space in a scene. To detect the shot boundaries, we use two popular frame difference metrics, the histogram difference metric (HDM) and the spatial difference metric (SDM), to measure the dissimilarity between the adjoining frame pair [23], [31], [38].

Given a frame size of $M \times N$, the spatial difference metric SDM is defined as

$$D_S(t) = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} |I_t(i,j) - I_{t+1}(i,j)| \tag{6}$$

where, $I_t(i,j)$ denotes the intensity of a pixel at location (i,j) in the tth frame.

The histogram difference metric HDM is defined as

$$D_H(t) = \frac{1}{M \times N} \sum_{k=1}^{L} |H_t(k) - H_{t+1}(k)| \tag{7}$$

where $H_t(k)$ denotes the gray-level or color histogram for the tth frame, and k is one of the L possible colors or gray-levels.

After computing the frame difference metrics from the video data, all frame pairs are mapped to a point-set in the feature space F_D spanned by the SDM and HDM metrics

$$F_D = \{F_D(t) = (D_S(t), D_H(t)),\ t = 1, 2, \ldots, T\}. \tag{8}$$

Due to the significant content change at the shot boundaries, the frame pair at a shot boundary usually has much larger SDM

and HDM features than regular frame pairs. However, this is not always the case, since sometimes a relatively fast camera movement may produce a sequence of large F_D values. Because we focus on detecting abrupt camera shot boundaries, the F_D values at the abrupt shot boundaries are usually local maxima points. In order to detect these local maxima points more reliably, we propose a new set of features called differential SDM and HDM, i.e., the difference between the original feature sequence and a median filtered feature sequence

$$DF_D = \{|F_D(t) - \text{Median}(F_D(t))|,\ t = 1, 2, \ldots, T\}. \quad (9)$$

Using the FCNN classifier, we can then classify the feature points into two categories corresponding to the shot-boundary frame pairs and the nonshot-boundary frame pairs. For reason of computational cost, only a portion of the input data are used as training samples. After the FCNN reaches convergence, the weight vectors are fixed and the trained network is used to compute membership values for all samples.

C. Caption (Dis)Appearance Detection

Following the shot boundary detection, we can detect the caption transition frames within each shot, since there is relatively small change in the scene between adjacent frames within each shot unless caption text changes. However, because the caption regions occupy only a small section of a video scene, they do not introduce significant changes in the spatial and histogram difference metrics (otherwise, a caption transition would have been classified as a shot boundary). Therefore we cannot use the SDM and HDM to detect caption transitions.

Instead, we propose a new metric, quantized spatial difference density (QSDD), to detect the caption transition frame. We first observe that the reason that SDM fails to detect caption transitions is that the small movement of a scene between adjacent frames produces many residual edge pixels at object boundaries in the difference image. A direct summation of these edge pixels may result in a value higher than that caused by caption transition. Fortunately, most of these edge pixels are sparsely distributed, while the residual pixels produced by the caption are highly concentrated because of the dense stroke pattern of Chinese characters. So we compute a feature that can measure the residual pixel density distribution.

The QSDD metric is defined by a two-step binarization of the difference image between a pair of adjacent frames. Let $I_t(i,j)$ denotes the pixel intensity at location (i,j) in the video frame f_t. First, a pair-wise pixel comparison is performed on an adjacent frame pair

$$\triangle I_t(i,j) = \begin{cases} 1 & \text{if } |I_t(i,j) - I_{t+1}(i,j)| > \theta_1 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where θ_1 is a threshold estimated from caption brightness. Then the obtained binary map is uniformly partitioned into K blocks $B_k, k = 1, 2, \ldots, K$ with a size of $b \times b$. Each block is labeled as significant change or no significant change according to the formula

$$\beta_t(k) = \begin{cases} 1 & \text{if } \sum_{(i,j)\in B_k} \triangle I_t(i,j) > \theta_2 \cdot b^2 \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

where θ_2 is called the filling rate. Then the QSDD metric simply counts the number of blocks with significant change, i.e., $\text{QSDD}(t) = \sum_{k=1}^{K} \beta_t(k)$. Since the caption residual pixels are closely distributed in the difference image, they tend to produce more blocks with significant changes. Thus the difference image at a caption transition has a high QSDD value.

Fig. 3 shows a video sequence example with eight caption transitions. We plot the SDM and HDM metrics in Fig. 3(a) and (b) respectively, where the dashed lines denote the shot boundaries. As we can see, it is difficult to detect caption transitions based on the SDM or HDM metrics. On the other hand, the QSDD plot in Fig. 3(c) clearly shows the caption transition locations. Fig. 4(a) and (b) displays two frame pairs and their differences at the caption transition locations marked as (1) and (2) in Fig. 3(c). We can easily see the caption regions in the difference image.

Due to a fast camera movement or a fast moving object in the scene, the QSDD can sometimes be fairly high. Fig. 4(c) and (d) demonstrates a couple of such examples. In Fig. 4(c), a large camera movement produces a large number of strong edges in the difference image of two adjacent frames, resulting in a high QSDD marked as (3) in Fig. 3(c). In Fig. 4(d), a fast moving car in the scene also gives a high QSDD which is marked as (4) in Fig. 3(c). However, unlike the caption transition which has a single large-QSDD value, these frames usually produce a continuous sequence of large QSDD values. Similar to our approach in shot boundary detection, to distinguish these continuous large-QSDD values from the caption transitions, we compute the differential QSDD, i.e., the difference between the QSDD and a median filtered QSDD. The result for the previous example is shown in Fig. 3(d). Now the false large QSDD sequences are significantly reduced, and the caption transitions can easily be located in the differential QSDD with a proper threshold β_0. Finally, to check whether there is a caption transition at a shot boundary, we compare the caption regions at the two caption transitions right before and after the shot boundary. If they are the same, then no caption transition happens. Otherwise, we consider them as two different captions.

D. Caption Region Location

At each caption transition, the difference image is computed by $\triangle I(i,j) = |I_t(i,j) - I_{t-1}(i,j)|$. As we can see in Fig. 4, the caption regions in the difference image stand out more clearly than in the original images. Therefore we choose to locate the caption position in the difference image. Since other objects in a scene may also produce boundary edges in the difference image, the binary frame difference measure used to compute QSDD is not reliable for locating the caption position. To this end, we develop a set of new features.

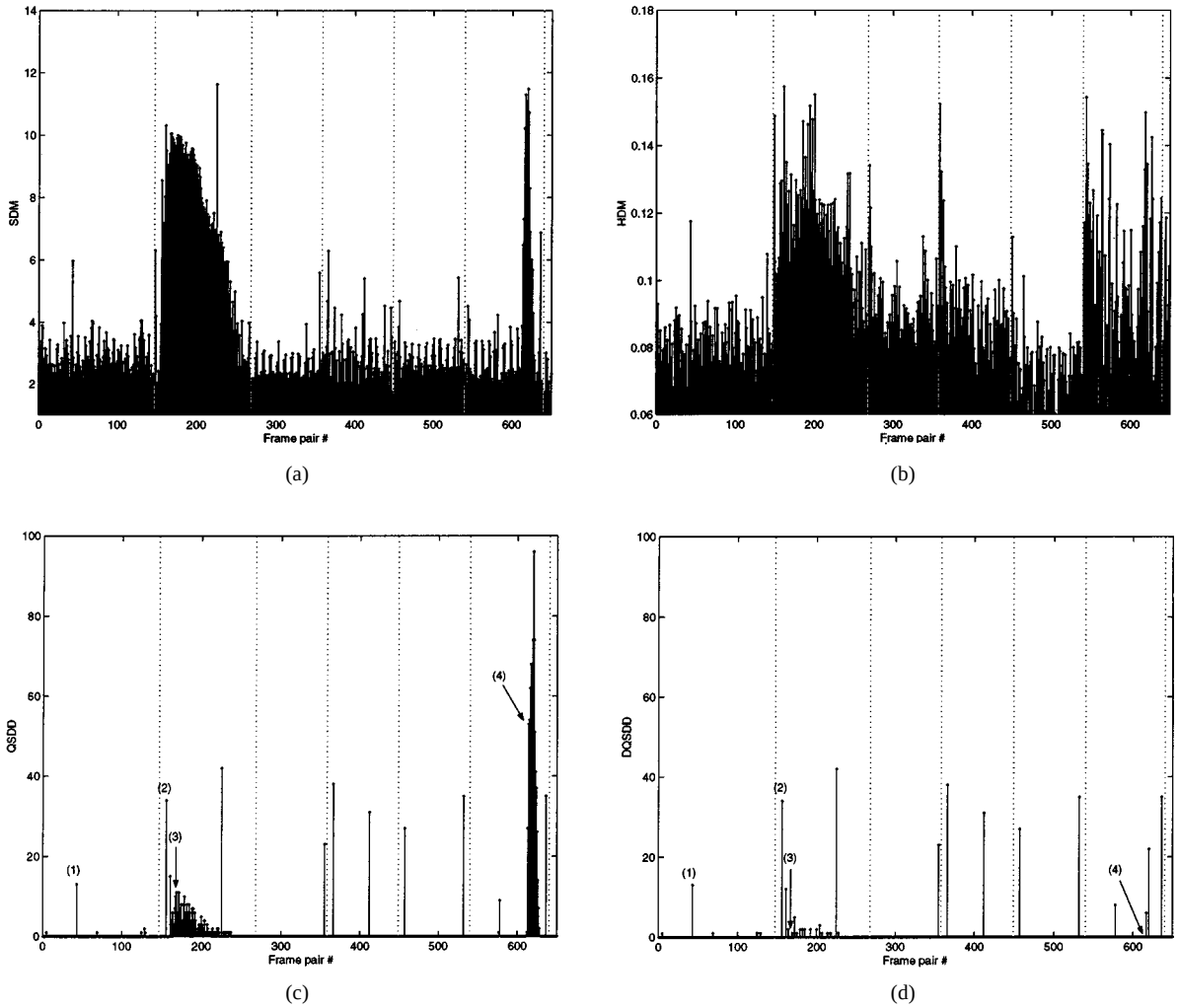


Fig. 3. Plots of frame difference metrics for a video sequence (a) SDM. (b) HDM. (c) QSDD. (d) Differential QSDD.

Since the Chinese characters are composed of four types of primitive strokes, i.e., horizontal, vertical, up-right-slanting and up-left-slanting strokes, the caption region contains rich edge information in the four directions. We use a set of four simple edge filters $M = \{M_1, \ldots, M_4\}$

$$M_1 = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix} \quad M_2 = \begin{pmatrix} 1 & -1 \\ 1 & -1 \end{pmatrix}$$
$$M_3 = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad M_4 = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

to filter the difference image. Then each filtered output is divided into K blocks, B_t^k, $k = 1, 2, \ldots, K$ with a block size of $b \times b$. We compute the means and variances of each block and its eight neighbors in all the four filtered images to construct a feature vector f_k with $m(= 2 \times 9 \times 4 = 72)$ elements.

To reduce the feature vector dimension and the correlation among the feature elements we use the principal component analysis (PCA) method [7]. Fig. 5 shows the scatter plots of the top six features obtained through the PCA method for a set of caption blocks and noncaption blocks. The caption block feature points are concentrating in the feature space center surrounded by the widely distributed noncaption block features. Using the FCNN classifier, we can separate the two classes quite easily. Several caption region location results are illustrated in Fig. 6.

III. Character Enhancement and Recognition

Although the caption region is located, it cannot be directly recognized by a regular OCR package due to the complex background and the low character resolution. Character enhancement and binarization is needed before character recognition.

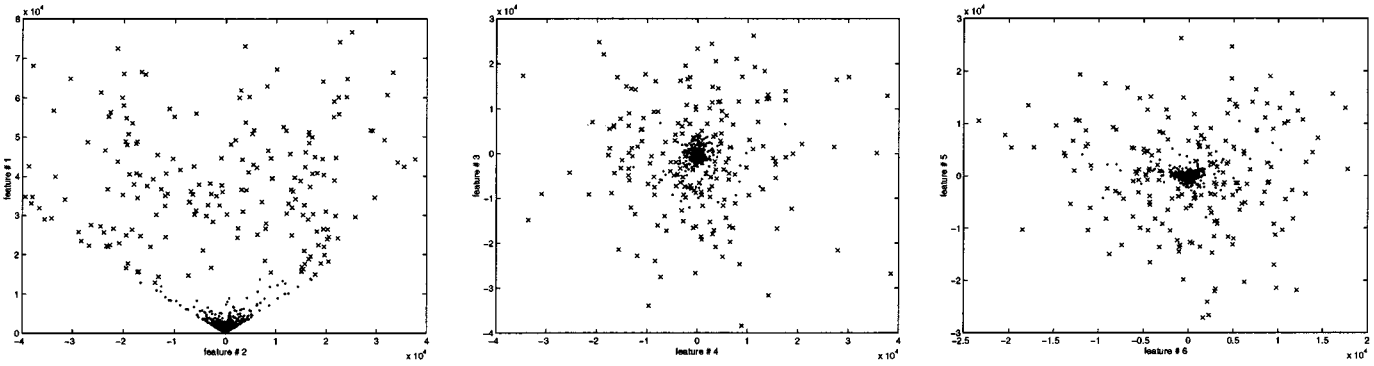
A. Caption Enhancement

Using the caption transitions detected in Section II, we can partition the video sequence into frame clusters. Within each cluster, the caption text stays the same. Multi-frame integration based caption enhancement techniques, such as the minimum pixel search [27] and frame averaging method [18], can be employed to reduce the influence of the complex background. For a frame cluster C_i with caption, each frame $f_t \in C_i$ contains a caption region $\gamma_i(f_t)$. The minimum pixel search method enhances the caption image by

$$\hat{\gamma}_i = \min_{f_t \in C_i} \gamma_i(f_t). \tag{12}$$

(a) (b) (c) (d)

Fig. 4. Frame difference examples. Adjacent frame pairs are shown in the first two rows. Their difference images are shown in the third row. (a) and (b) are caption transitions, (c) has a fast camera movement and (d) has a fast moving car in the scene. The locations of (a)–(d) in the video sequence are marked as (1)–(4) in Fig. 3.



(a) (b) (c)

Fig. 5. Scatter plots of the top six features extracted by applying the principal component analysis on the original features. "•" denotes the caption region samples and "×" denotes the noncaption region samples.

For the stable character strokes, the output should stay relatively the same, while the changing background is minimized. Similarly, the frame-average method enhances the caption image by

$$\bar{\gamma}_i = \frac{1}{|C_i|} \sum_{f_t \in C_i} \gamma_i(f_t) \tag{13}$$

where $|C_i|$ denotes the number of frames in C_i. Since the fluctuation of the background is greater than the caption, the averaging operation should reduce the variation of the complex background.

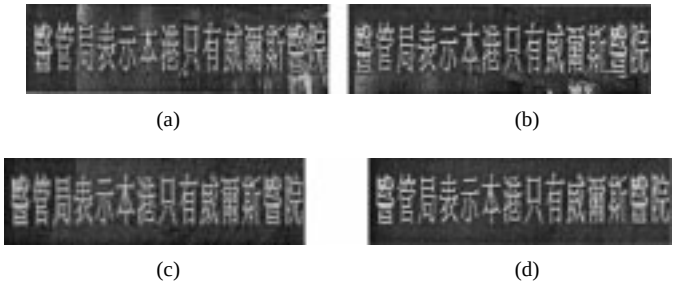
Fig. 7 shows an example of caption enhancement by using the above two methods. The complex background is reduced greatly by integrating the multiple frame images. The key in implementing the two methods is the registration of a maximum number of caption blocks with the same text string. Our caption (dis)appearance detection and caption region location algorithms provide just that.

B. Character Segmentation

For the enhanced caption images, it is still difficult to separate the texts from the background by a simple threshold. Some multiple threshold methods have been proposed [2], [18], [35], [36], including adaptive thresholding and floating three threshold method. Given the unlimited background variations across the caption region, it is difficult to deal with different variations at different parts of a caption region at the same time. Here, we propose a new approach in which the caption region is first separated into individual characters, then the binarization is performed on each individual character image. Since the background of each character is relatively simple, a regular thresholding method can be used.

(a) (b) (c) (d)

Fig. 6. Caption region location examples. (a) Difference images of the caption-transition frame pairs. (b) Classified label maps. (c) Segmented text regions in the difference images. (d) Caption bounding boxes in the original frames.



(a) (b) (c) (d)

Fig. 7. Caption enhancement examples. (a) and (b) Captions in the twentieth and sixty-fourth frame of a video sequence. (c) Enhanced caption using the minimum-pixel-search scheme. (d) Enhanced caption using the frame averaging scheme.

(a) (b) (c) (d)

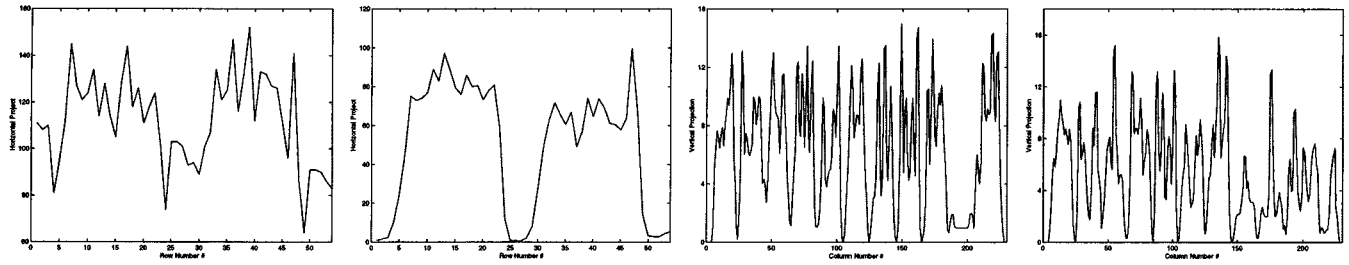
Fig. 8. Character segmentation examples. (a) Caption region. (b) Binary image. (c) Smoothed contour image. (d) Located individual characters.

The individual character extraction from a caption image consists of two steps: text-line separation followed by individual character separation in a text line. In the traditional document OCR, these two steps are carried out by horizontal and vertical projections. However, these simple projections are not suitable for video frames due to the complex background.

To reduce the influence of complex backgrounds, we first binarize the caption region using the intensity mean as a simple threshold. As shown by the example in Fig. 8(b), a couple of large patches of backgrounds cannot be separated from the characters. To further remove the background, we use a set of mathematical morphology operations to detect the contour of the characters. Let R_c denote the binary caption region, the contour map of the caption text is extracted by

$$\delta(R_c) = R_c - R_c \circ S = R_c - (R_c \ominus S) \oplus S \qquad (14)$$

where S is an isotropic structuring element, "$\circ$" is the opening operation which is defined by an erosion operator "$\ominus$" followed by a dilation operator "$\oplus$." We then apply a smoothing filter on the contour map to eliminate extraneous fragments and to connect broken character segments. Fig. 8(c) shows the obtained contour map from the binary image (b). Now we can use a horizontal projection to separate the text lines. In Fig. 9(a) and (b) are the horizontal projections of the binary image and contour map in Fig. 8(b) and (c), respectively.



(a) (b) (c) (d)

Fig. 9. Caption projection examples. (a) The horizontal projection of the binary image in Fig. 8(b). (b) The horizontal projection of the smoothed contour image in Fig. 8(c). (c) The vertical projection of the first caption line in Fig. 8(c) and (d) the vertical projection of the second caption line in Fig. 8(c).

Clearly, the caption lines can be separated easily from the smoothed contour image instead of the binary image.

For each obtained text line, a vertical projection of the contour map can be used to locate the individual characters. Shown in Fig. 9(c) and (d) are the vertical projections of the two caption lines in Fig. 8(c). Even though most of the characters can be separated by a fixed threshold, it is difficult to select a proper threshold in order to locate all the characters. A high threshold will lead to character loss, while a low threshold will lead to character merging. Our strategy is first to guarantee a low character loss rate, then to re-segment the merged characters using three heuristic rules derived from the square shape of Chinese characters.

1) If the width of a segment W matches its height H, i.e., $0.6H \leq W \leq 1.4H$, it is declared as a character image.
2) If the width of a segment $W < 0.6H$, and if the follow up character is closer than a preset distance, then merge the segment into the next character region. Otherwise, discard the segment.
3) If the width of a segment $W > 1.4H$, then use the average character width that satisfy rule (1) to repartition this segment.

Fig. 8(d) shows the located boundaries of each character for the caption example in (a).

C. Binarization and Recognition

Although we have obtained the individual characters, they cannot be put into an OCR classifier directly, since the extreme low resolution is insufficient for recognition and the character is still blended with a complex background. Before separating the character from the background, we first increase the resolution of the character image by a factor of eight through interpolation. Even though this does not add new information to the gray scale image, it does help to smooth the character boundary in the binary image.

We show the character binarization steps through an example character shown in Fig. 10. Fig. 10(a) gives a typical character in a nonuniform background. Using a spline interpolation function, we obtain the image with higher resolution in (b). We then binarize the interpolated image with a fixed threshold to get the binary image in (c). Apparently, some background regions still remain in the binary image. Fortunately, a bright character on a bright background always has a black profile around it to help the audience to read the character. Based on this observation, a region connectivity analysis scheme is proposed to eliminate the remaining background residues.

(a) (b) (c) (d)

Fig. 10. An example of character binarization. (a) The original character. (b) The interpolated character. (c) The binarized character. (d) The clear character after filtering.

TABLE I
EXPERIMENTAL RESULTS OF CAPTION DETECTION

	Hits	Misses	False alarms	Precision	Recall
Shot boundary detection	2,954	59	33	98.90%	98.04%
Caption transition detection	3,445	28	321	91.47%	99.19%
Caption line location	914	4	28	97.44%	99.56%

TABLE II
COMPARISON OF THE CHINESE OCR RESULTS BETWEEN THE ORIGINAL CHARACTER IMAGES AND THE PROCESSED BINARY CHARACTERS

Candidate number	With processing	Without processing
1	85.82%	13.30%
2	90.79%	17.11%
3	92.10%	19.74%

Due to background noise, the character often connects the background residues with small bridges even though there is a black profile around the character. To break the bridges across the character and the background, we adopt a set of morphological processing techniques, the opening operation and H-break operation. The opening eliminates small noise particles, but preserves the global shape of the objects. The H-break operation mainly eliminates the H-connected pixels. Both of them are effective in cutting off the connection between the character and the background. After the morphological operations, we label all the connected-components in the binary image and remove components that are too small in size. If a connected component connects the periphery of the character image, it is also declared as background and filtered out. Through these processes, we finally obtain the binary character with clear background shown in Fig. 10(d).

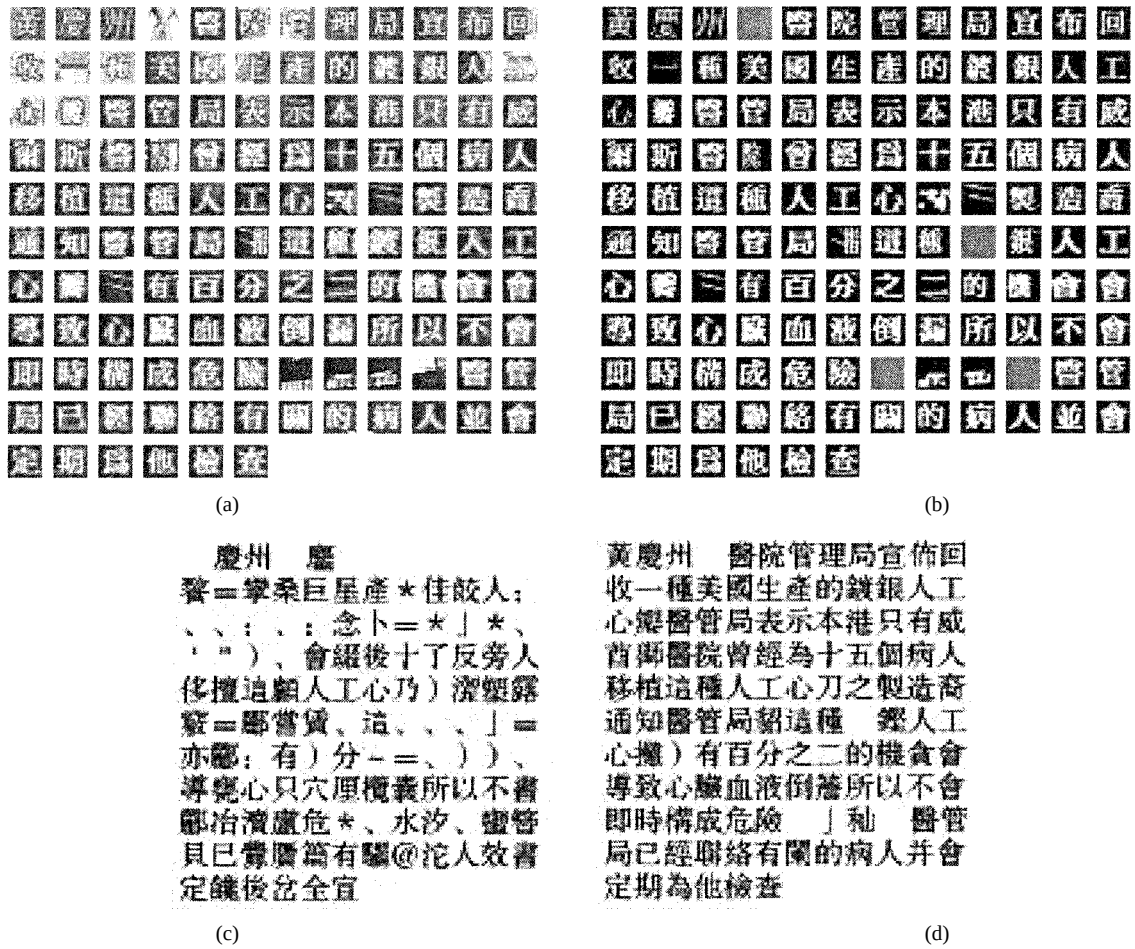


Fig. 11. Video caption OCR example. (a) A set of characters extracted from a news video story. (b) The processed characters. (c) The recognition results using (a). (d) The recognition results using (b).

Since there exist many successful commercial Chinese OCR packages, we do not intend to implement a new one. In our system, we use the OCR package TH-OCR LV for final character recognition. Our goal is to test whether our character extraction method can efficiently obtain binary characters with enough clarity for regular printed optical character recognition.

IV. Experimental Results

We evaluate the proposed caption detection and recognition system using the TVB and Asian TV news programs in Hong Kong. Video data are encoded in MPEG-1 format at a resolution of 288×352. Twelve 30-min programs are employed in the experiments.

A. Caption Detection Experiments

We first evaluate the three components of the caption detection module. For the 12 news videos, there are 3013 shot boundaries. The FCNN classifier detects 2954 real shot boundaries with 33 false alarms. While in the process of caption (dis)appearance detection, we need to make a tradeoff between the precision and the recall. In the experiments, we set the network learning parameter $\alpha_0 = 0.1$, and the maximum number of learning iteration $T = 500$. We empirically set $\theta_1 = 40$, $\theta_2 = 0.6$, and select a low threshold $\beta_0 = 12$ to ensure a high recall rate rather than a high precision rate. We hope to detect as many caption regions as possible with a reasonable accuracy, then remove most of the false alarms at later stages. Since the overall missing rate can only increase through each step, but the false alarm rate can be reduced through later processing, we always emphasis recall over precision in the first few steps of the system.

In the caption (dis)appearance detection experiment, we use 300 video shots containing 3473 caption transitions. Our caption (dis)appearance detection scheme detects 3766 caption transitions and misses 28 real caption transitions. The missed caption transitions are mainly due to the gradual appearance or disappearance of captions, which is difficult to detect by using our method. Fortunately, unlike movie programs, news programs seldom use this type of special effect captions. The false detections are caused by the moving scene or objects which are not completely removed by the differential QSDD.

In order to test the caption region location algorithm, we select 600 difference images at the caption transition locations. There are totally 918 caption lines in these difference images. Our program detects 914 real caption lines and produces 28 false alarms. Thus, for both the caption transition detection and caption line location, we achieve a recall rate of over 99%. There are very few caption text lines missed by our approach. All the experimental results for the caption detection module are summarized in Table I.

B. Character Recognition Experiment

For the extracted characters, we first enlarge their image size by a factor of eight using spline interpolation. Then we perform the binarization and morphological processing to obtain the clean binary characters for final recognition. The recognition performance is compared with the recognition results using the original character images. In the experiments, the threshold for binarizing the interpolated characters is selected as 0.8 times of the maximal character intensity. The minimum size of a connected component is set to 50. The results for five news stories are shown in Table II. In this table, we compute the correct recognition rate by using the first candidate, the first two candidates, and the first three candidates, respectively. The results show that our processing scheme improves the correct recognition rate by nearly five folds.

As an example, Fig. 11 illustrates the caption OCR result of a news story. The extracted characters are shown in Fig. 11(a). Fig. 11(b) gives the result of the binarization and morphological processing. We can see that three false characters extracted in the first step are excluded, and one character is lost in this step. The OCR results for the characters in Fig. 11(a) and (b) are shown in Fig. 11(c) and (d), respectively. Only 21 characters are correctly recognized for the original images, while 108 characters are correctly recognized for the processed images.

V. Conclusion

We have developed an automatic video-caption detection and recognition system. Using an unsupervised FCNN classifier to focus on caption transition frames, the system detects video captions with high precision and efficiency. Furthermore, aided by a set of novel character segmentation and binarization methods, we improve the Chinese video OCR accuracy from 13% to 86% for a set of Chinese news videos. As the first attempt on Chinese video caption recognition, these results are very encouraging.

The caption detection module cannot deal with special effect moving-captions. However, since these types of captions are used far less frequently than regular captions, we can afford to employ more complex caption detection and tracing algorithms, such as the ones developed in [19], to detect them. The character segmentation and binarization module in our system can be applied to not only video captions but also low-quality document character recognition. In addition, most of the algorithms developed in this paper can be easily applied to caption extraction in any video programs of other languages with minor modifications.

Acknowledgment

The authors wish to thank Dr. Q. Yang for many constructive comments and the Hong Kong TVB and Asian TV news stations for the news videos.

References

[1] L. Agnihotri and N. Dimitrova, "Text detection for video analysis," in Proc. Workshop Content-Based Access Image Video Libraries, Conjunction CVPR, Fort Collins, CO, June 1999.

[2] Y. Ariki, K. Matsuura, and S. Takao, "Telop and flip frame detection and character extraction from TV news articles," in *Proc. 5th Int. Conf. Document Anal. Recognition*, 1999, pp. 701–704.

[3] Y. A. Aslandogan and C. T. Yu, "Techniques and systems for image and video retrieval," *IEEE Trans. Knowledeg Data Eng.*, vol. 11, pp. 56–63, 1999.

[4] Y. Avrithis, N. Tsapatsoulis, and S. Kollias, "Broadcast news parsing using visual cues: A robust face detection approach," in *Proc. 2000 IEEE Int. Conf. Multimedia Expo.*, vol. 3, 2000, pp. 1469–1472.

[5] Y. Cui and Q. Huang, "Character extraction of license plates from video," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 1997, pp. 502–507.

[6] K. Etemad, D. Doermann, and R. Chellapa, "Multiscale segmentation of unstructured document pages using soft decision integration," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, pp. 92–96, 1997.

[7] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1990.

[8] C. Garcia and X. Apostolidis, "Text detection and segmentation in complex color images," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 4, 2000, pp. 2326–2329.

[9] U. Gargi, D. Crandall, S. Antani, T. Gandhi, R. Keener, and R. Kasturi, "A system for automatic text detection in video," in *Proc. ICDAR'99*, 1999, pp. 29–32.

[10] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray, "Automated generation of news content hierarchy by integrating audio, video, and text information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 6, 1999, pp. 3025–3028.

[11] A. K. Jain and B. Yu, "Automatic text location in images and video frames," *Pattern Recognition*, vol. 31, no. 12, pp. 2055–2076, 1998.

[12] K. Y. Jeong, K. Jung, E. Y. Kim, and H. J. Kim, "Neural network based text location for news video indexing," in *Proc. Int. Conf. Image Processing*, vol. 3, 1999, pp. 319–323.

[13] H. K. Kim, "Efficient automatic text location method and content-based indexing and structuring of video database," *J. Visual Commun. Image Representation*, vol. 7, no. 4, pp. 336–344, 1996.

[14] S. K. Kim, D. W. Kim, and H. J. Kim, "Recognition of vehicle license plate using a genetic algorithm based segmentation," in *Proc. ICIP*, 1996, pp. 661–664.

[15] S. Kurakake, H. Kuwano, and K. Odaka, "Recognition and visual feature matching of text region in video for conceptual indexing," in *Proc. SPIE Storage Retrieval Image Video Databases 3022*, 1997, pp. 368–379.

[16] H. Kuwano, Y. Taniguchi, H. Arai, M. Mori, S. Kurakake, and H. Kojima, "Telop-on-demand: Video structuring and retrieval based on text recognition," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2, 2000, pp. 759–762.

[17] C. M. Lee and A. Kankanhalli, "Automatic extraction of characters in complex scene images," *Int. J. Pattern Recognition Artificial Intell.*, vol. 9, no. 1, pp. 67–82, 1995.

[18] H. P. Li, D. Doemann, and O. Kia, "Text extraction, enhancement and OCR in digital video," in *Proc. 3rd IAPR Workshop*, Nagoya, Japan, 1998, pp. 363–377.

[19] ——, "Automatic text detection and tracking in digital video," *IEEE Trans. Image Processing*, vol. 9, pp. 147–156, 2000.

[20] R. Lienhart and F. Stuber, "Automatic text recognition in digital videos," in *Proc. SPIE Image Video Processing IV 2666*, 1996, pp. 180–188.

[21] M. Maybury, A. Merlino, and J. Rayson, "Segmentation, content extraction and visualization of broadcast news video using multistream analysis," in *Proc. AAAI Spring Symp.*, Stanford, CA, 1997, pp. 1–12.

[22] S. Mitra and S. K. Pal, "Self-organizing neural network as a fuzzy classifier," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, pp. 385–399, 1994.

[23] A. Nagasaka and Y. Tanaka, "Automatic video indexing and full video search for object appearances," in *IFIP Trans., Visual Database Syst. II*, E. Knuth and L. M. Wegner, Eds. New York: Elsevier, 1992.

[24] J. Ohya, A. Shio, and S. Akamatsu, "Recognizing characters in scene images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 16, pp. 214–220, 1994.

[25] N. R. Pal, C. Bezdek, and E. C.-K. Tsao, "Generalized clustering networks and Kohonen's self-organizing scheme," *IEEE Trans. Neural Networks*, vol. 4, pp. 549–557, July 1993.

[26] G. Picciolio, E. De Michieli, P. Parodi, and M. Campani, "Robust method for road sign detection and recognition," *Image Vision Comput.*, vol. 14, pp. 35–46, 1996.

[27] T. Sato, T. Kanade, E. K. Kughes, M. A. Smith, and S. Satoh, "Video OCR: Indexing digital news libraries by recognition of superimposed captions," *ACM Multimedia Syst. (Special Issue on Video Libraries)*, vol. 7, no. 5, pp. 385–395, 1999.

[28] J. C. Shim, C. Dorai, and R. Bolle, "Automatic text extraction from video for content-based annotation and retrieval," in *Proc. Int. Conf. Pattern Recognition*, 1998, pp. 618–620.
[29] M. A. Smith and T. Kanada, "Video skimming and characterization through the combination of image and language understanding technique," in *Proc. IEEE Conf. Comput. Vision Pattern Recognition*, 1997, pp. 775–781.
[30] K. Sobottka, H. Bunke, and H. Kronenberg, "Identification of text on colored book and journal covers," in *Proc. ICDAR'99*, 1999, pp. 57–62.
[31] H. Ueda, T. Miyatake, and S. Yonizawa, "IMPACT: An interactive natural-motion-picture dedicated multimedia authoring system," in *Proc. CHI'91*, New Orleans, LA, 1991, pp. 343–350.
[32] H. Wactlar, T. Kanade, M. Smith, and S. Stevens, "Intelligent access to digital video: Informedia project," *IEEE Trans. Comput.*, vol. 10, pp. 46–52, 1996.
[33] A. Wernicke and R. Lienhart, "On the segmentation of text in videos," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 3, 2000, pp. 1511–1514.
[34] E. K. Wong and M. Chen, "A robust algorithm for text extraction in color video," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2, 2000, pp. 797–800.
[35] V. Wu, R. Manmatha, and E. M. Riseman, "Finding text in images," in *Proc. 2nd ACM Int. Conf. Digital Libraries*, 1997.
[36] ——, "TextFinder: An automatic system to detect and recognize text in images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 1224–1229, Nov. 1999.
[37] D. Zhang and S. K. Pal, "A fuzzy clustering neural networks (FCN's) sytem design methodology," *IEEE Trans. Neural Networks*, vol. 11, pp. 1174–1177, Sept. 2000.
[38] H. J. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic partitioning of video," *Multimedia Syst.*, vol. 1, pp. 10–28, 1993.
[39] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," *Pattern Recognition*, vol. 28, no. 10, pp. 1523–1535, 1995.
[40] J. Zhou, D. Lopresti, and T. Tasdizen, "Finding text in color images," in *Proc. SPIE, Document Recognition V*, 1998, pp. 130–140.

Xiaoou Tang (S'93–M'96) received the B.S. degree in 1990 from the University of Science and Technology of China, Hefei, China, and the M.S. degree in 1991 from the University of Rochester, Rochester, NY. He received the Ph.D. degree in 1996 from the Massachusetts Institute of Technology (MIT), Cambridge.

He is currently an Assistant Professor in the Department of Information Engineering of the Chinese University of Hong Kong. His research interests include video processing and pattern recognition.

Xinbo Gao (M'02) received the B.S., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China, in 1994, 1997, and 1999, respectively.

Since 1999, he has been with the School of Electronic Engineering, Xidian University, China, where he is currently an Associate Professor at the Department of Electronic Engineering. From 2000 to 2001, he was with the Department of Information Engineering, the Chinese University, Hong Kong as a Research Associate. His research interests include content-based video analysis and representation, image understanding, pattern recognition, and artificial intelligence.

Jianzhuang Liu received the B.E. degree from Nanjing Institute of Posts and Telecommunications, China, in 1983, the M.E. degree from Beijing University of Posts and Telecommunications, China, in 1987, and the Ph.D. degree from the Chinese University of Hong Kong, China, in 1997.

From 1987 to 1994, he was a Member of the Academic Staff in the Department of Electronic Engineering, Xidian University, China. From August 1998 to August 2000, he was a Research Fellow at the School of Mechanical and Production Engineering, Nanyang Technological University, Singapore. He is now a Postdoctoral Fellow at the Chinese University of Hong Kong, China. His research interests include image processing, computer vision, pattern recognition, computer graphics, and artificial intelligence.

Hongjiang Zhang (S'90–M'91–SM'97) received the B.S. degree from Zhengzhou University, China and the Ph.D. degree from the Technical University of Denmark, both in electrical engineering, in 1982 and 1991, respectively.

From 1992 to1995, he was with the Institute of Systems Science, National University of Singapore, where he led several projects in video and image content analysis and retrieval and computer vision. He also worked at MIT Media Lab in 1994 as a Visiting Researcher. From 1995 to 1999, he was a Research Manager at Hewlett-Packard Labs, where he was responsible for research and technology transfers in the areas of multimedia management; intelligent image processing and Internet media. In 1999, he joined Microsoft Research Asia, where he is currently a Senior Researcher and Assistant Managing Director, mainly in charge of media computing and information processing research. He has authored three books, more than 170 referred papers and book chapters, seven special issues of international journals in multimedia processing, content-based media retrieval, and Internet media, as well as numerous patents or pending applications.

Dr. Zhang is a Member of ACM. He currently serves on the editorial boards of five professional journals and a dozen committees of international conferences.