Pattern Anal Applic (2007) 10:69–81
DOI 10.1007/s10044-006-0051-9

THEORETICAL ADVANCES

Extracting content from instructional videos by statistical modelling and classification

Chekuri Choudary · Tiecheng Liu

Received: 8 January 2006 / Accepted: 6 September 2006 / Published online: 15 November 2006

Abstract This paper presents a robust approach to extracting content from instructional videos for handwritten recognition, indexing and retrieval, and other e-learning applications. For the instructional videos of chalkboard presentations, retrieving the handwritten content (e.g., characters, drawings, figures) on boards is the first and prerequisite step towards further exploration of instructional video content. However, content extraction in instructional videos is still challenging due to video noise, non-uniformity of the color in board regions, light condition changes in a video session, camera movements, and unavoidable occlusions by instructors. To solve this problem, we first segment video frames into multiple regions and estimate the parameters of the board regions based on statistical analysis of the pixels in dominant regions. Then we accurately separate the board regions from irrelevant regions using a probabilistic classifier. Finally, we combine top-hat morphological processing with a gradient-based adaptive thresholding technique to retrieve content pixels from the board regions. Evaluation of the content extraction results on four full-length instructional videos shows the high performance of the proposed method. The extraction of content text facilitates the research on full exploitation of instructional videos, such as content enhancement, indexing, and retrieval.

Keywords Instructional video analysis · Video text detection · E-learning

C. Choudary (✉) · T. Liu
Department of Computer Science and Engineering,
University of South Carolina, Columbia, SC 29208, USA
e-mail: choudary@cse.sc.edu

1 Introduction

Analyzing the content of instructional videos is of great importance for advanced learning technologies [6]. In instructional videos, the video content (i.e., written or printed text and figures) is essential for indexing, retrieval, and other e-learning applications. Therefore, it is critical to accurately retrieve the content text and figures from these videos. In a real teaching environment, instructors may use a variety of narrative forms, such as electronic slides (e.g., PowerPoint), handwritten slides, and board presentations. Nevertheless, the presentation using chalkboards remains one of the most adopted narrative forms, especially in math-intensive courses. Among these narrative forms, the content on electronic slides is directly accessible via computer files, and the content on paper or transparency slides can be easily retrieved using regular document analysis techniques. However, detecting and extracting the written content on boards is still challenging and is more difficult than content extraction from other media.

In this work, we focus on the board presentation videos and present our approach to accurately extract content from these videos. Throughout this paper, we always refer to instructional videos as the board presentation videos and refer to *video content* as the characters, drawings, and figures on chalkboards written by instructors. The problem of extracting instructional video content is a classification problem in nature. For each video frame, we want to classify image pixels into three categories: the *content pixels*, the *board pixels*, and the *irrelevant pixels*. The content pixels are the handwritten chalk pixels on boards; the board pixels are the non-chalk pixels in the board regions; the irrelevant pixels are those that belong to neither of the two previous categories. The work presented in this paper can be used as the first step that facilitates the research on handwritten recognition, event detection [30], structuring [7, 19, 23], and indexing and retrieval of instructional videos [1, 8, 22].

1.1 Related works and the challenges of instructional videos

Extracting content text from instructional videos is challenging and is drastically different from text detection in other domains, in the following three aspects.

First, most existing works [3, 4, 9, 14, 17, 24, 31] on video text detection are developed mainly for optical characters that have fixed fonts and clear text lines, such as the enclosed captions in videos. However, they are not tuned to handwritten characters and figures that vary significantly in size and edge density. In [31], Tang et al. proposed a system for extracting overlaid captions in videos. Each pixel is classified as a caption pixel or a background pixel based on the variation of the pixel intensities in the video shot. Overlaid captions are assumed to have fairly constant values, whereas background pixels are supposed to show a lot of variation in the intensity values. This technique does not fit the application of extracting content text on boards because (1) the text and figures are developed dynamically in the instructional videos, and (2) the content is often occluded by the instructor.

Edge-based approaches [4, 9, 24] are commonly employed in video text detection. In [24], Ngo and Chan developed a coarse-to-fine approach to retrieve video text based on edges. Cai et al. [4] developed a system to extract the text embedded in complex backgrounds. The edge-based features (such as edge strength, edge density, and the projection profiles of the edge map) are used to identify the horizontal and vertical text strings. In [9], the aspect ratios of edges and the existence of raising and falling edges are used to find the text regions in natural scenes. Gaussian mixture model is used to extract the text from these regions. The above methods are not quite suitable for detecting chalk pixels on boards because the handwritten characters, drawings, and figures do not show uniform edge density. In addition, unlike the overlaid video captions, the chalk pixels have weak contrast with the board background, making purely edge-based methods not feasible.

In [3, 14, 17], the bounding boxes of text lines are first detected. However, in instructional videos, it is difficult to retrieve the text bounding boxes due to the lack of clear text lines in the handwritten figures and characters. For electronic slide presentations, Liu et al. [18] developed a video segmentation method to partition lecture videos into video segments, each of which relates to one slide. Since the text can be directly retrieved from the electronic slides, there is no need to extract video text in electronic slide presentations. In [12], Ju et al. proposed a content analysis method for the videos of transparency slides. In [2], Mittal et al. provided a content-based compression technique for chalkboard videos. However, these works did not provide methods to extract content text. Ju et al. [12] focus on detecting and analyzing hand gestures in lecture videos. Ankush Mittal et al. [2] segment the video frames into board region, instructor, and background to apply different compression methods for the three objects. The segmentation of board region is based on horizontal edges and cannot be applied when there is camera motion and zooming.

Second, the text and figures on boards cannot be extracted by a trivial thresholding operation. In real classrooms, the board is not always “black”: it may have some dark colors like dark green or dark blue. In some light conditions, the board may even show light grey color. The luminance and the color of the board background vary temporally among video frames due to the change of light conditions. Within each frame, the pixels in board regions have significant variation in luminance due to the wear of the board and light reflection. During a class session, instructors may erase some chalk content, and the chalk dust left by erasures causes the erased regions to show high luminance value, making content extraction more difficult.

In [11], Heng and Tian developed a system for content enhancement of lecture videos. This system is based on simple thresholding of pixel intensity, thus it is prone to extract false content pixels due to image noise, chalk erasures, and variation of luminance in board regions. Although there is extensive research on text detection from printed or scanned documents [28] and camera captured images [15], the existing methods cannot be directly applied to instructional videos because they cannot handle the occlusions in video frames and the noise in board regions. In Sect. 5, we show the results of two existing text detection methods (Kittler’s method [13] and Niblack’s method [25]) on instructional videos, comparing to the results of our proposed method.

Third, in instructional videos, camera movements and occlusions are very common and unavoidable. The board content and the classroom activities are usually recorded non-intrusively by a camera (or cameras) mounted at the back of the classroom. Due to the limitations of the camera resolution and the dimension of the board, the cameras usually need to be zoomed-in to capture a portion of the board. Constrained by the nature of board presentations, instructors unavoidably occlude part of content text while they write on boards. In [26], Onishi et al. presented the research on detecting blackboard content based on background subtraction. A similar approach was employed in [29] for analyzing the handwritten content on white board. However, Onishi's work is restrictive because it needs the registration of a clear board background with no chalk content and occlusions. For the classrooms that have multiple board panels, the switching of panels makes simple image registration technique [26] degrade in performance. In [33], Wienecke et al. developed a system for automatically reading the white board content. This method relies on edge density to detect text blocks, and the extraction is a standard thresholding operation. This method does not consider the excessive occlusions and camera motions. In Fig. 1, we show sample instructional video frames which indicate the difficulties of content analysis in real classroom videos.

1.2 Our approach

The proposed approach for extracting instructional video content consists of three steps: preprocessing and parameter estimation, background separation, and content extraction. (1) In the preprocessing step, sampled instructional video frames are first segmented into multiple regions using an existing image segmentation method [5]. Then we estimate the average and the variance of the color of the board background, and the range of luminance by clustering and separating the outliers among the average colors of the dominant regions. (2) With the estimated parameters of the boards, the background separation step accurately separates the content regions from the irrelevant regions by clustering image regions. The results are further refined by topological analysis of the regions. (3) In the content extraction step, we binarize the content pixels of text and figures by performing the top-hat morphological processing followed by a gradient-based adaptive thresholding. These three processing steps are presented in Sects. 2, 3, and 4, separately. Section 5 shows the experimental results, and Sect. 6 gives a short conclusion of the work and future directions.

2 Preprocessing and parameter estimation

The first step of our approach segments video frames into regions and estimates the parameters of the board color and luminance. As instructional videos are highly redundant in visual content, it is not necessary to process all video frames. Our experiments indicate that at a fixed sampling rate of one frame per 150 frames (5 seconds), the selected frames still contain all the content text and figures. So in our experiments, we use this fixed sampling rate to retrieve a subset of frames for content analysis.

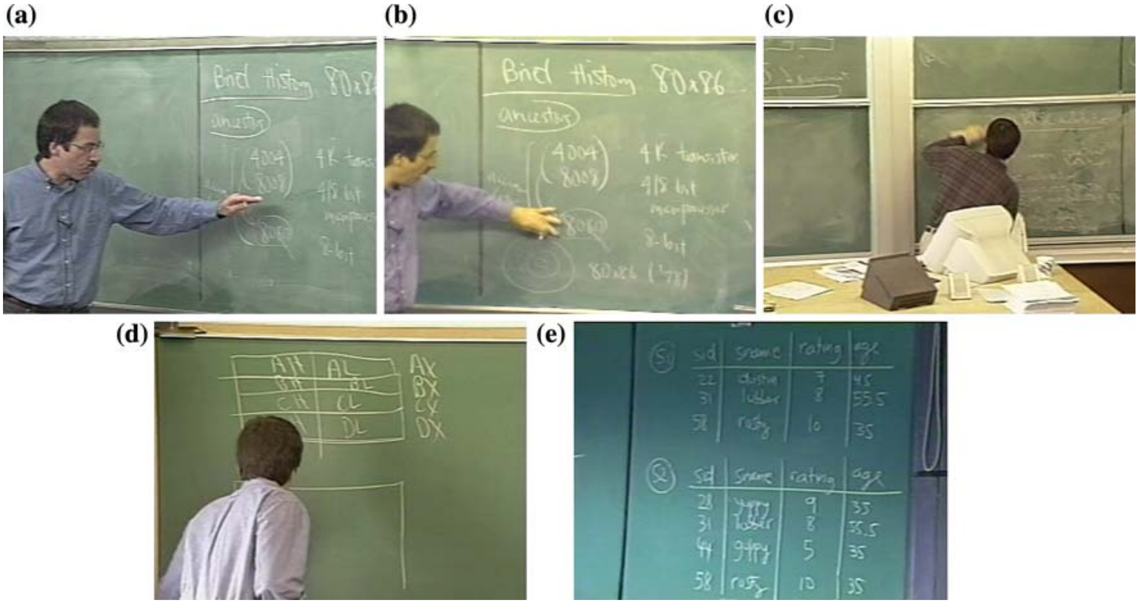

Fig. 1 Detecting and extracting instructional content recorded in real classrooms is challenging, as indicated in the above sample video frames: **a**, **b** color and luminance change of the pixels in board regions; **c** unclear content text because of camera zooming; **d** occlusions and shadows caused by instructors; **e** non-uniform illumination

The pixels of the board vary in color and luminance, both spatially across the board region and temporally during a video session. Therefore, it is critical to model the color and luminance distribution of the pixels of the board and to estimate the parameters of the model. In this paper, we define *board regions* as the regions that contain either the chalk pixels or the board background pixels; we define *irrelevant regions* as those regions other than the board regions. Note that one video frame may have multiple board regions because of multiple board panels and occlusions.

Segmentation can help the estimation of the parameters of the board model by partitioning an image into regions. However, existing segmentation methods [5, 21, 32] cannot directly separate the board regions from the irrelevant regions due to the following reasons: (1) Most segmentation methods are not developed for figure-ground segmentation and produce over-segmented regions for instructional video frames. Some chalk text lines and the regions of chalk erasures are usually mistakenly segmented as individual regions. (2) Although some methods [21, 32] are able to produce figure-ground segmentation, they are not tuned to this application, and the results are not accurate enough for content analysis. (3) When an image is separated into multiple regions, it is still not clear which regions constitute the board regions.

In our work, we first use the mean-shift segmentation developed by Comaniciu and Meer [5] to partition video frames into connected regions. The mean-shift method is selected because of its low computational complexity in clustering image pixels. In our experiments, we find that for a wide range of parameters (spatial bandwidth in the range of 6–18, color bandwidth in the range of 5–16, minimum region area in the range of 15–60), the mean-shift segmentation [5] produces over-segmented regions. To accurately separate the board regions, we need to estimate the parameters for the distribution of pixels in the board regions.

Two observations lead to our estimation of the board parameters. First, in most frames, the largest region segmented by mean-shift is a sub region of the board regions, so the average colors of the largest regions in video frames give a coarse estimation of the board color in the entire video sequence. However, due to occlusions and camera movements, there still exists the case where the largest region of one frame is an irrelevant region. Therefore, we need to analyze and remove the outliers for the estimation of the color variation of the board background. Second, due to light reflection, change of light condition, and chalk dust caused by erasures in instructional videos, the luminance of the board pixels varies in a wide range, yet the color components (a^* and b^* in $L^*a^*b^*$ color space) are usually concentrated in a small range. So it is essential to estimate the range of the luminance and the variance of the board colors.

Based on these two observations, we develop a simple clustering method for outlier separation and parameter estimation of the board background. In this paper, we use the $L^*a^*b^*$ color space because it is a perceptually uniform color space and it separates the luminance component from the color components. Figure 2 shows the distributions of the average colors of the largest regions in two instructional videos.

Suppose $\{f_1, f_2,..., f_n\}$ are the n sampled frames of an instructional video. Let $R^{(i)}$ be the largest region (i.e., dominant region) of the frame f_i segmented using the mean-shift method. We compute the average colors of the largest regions. Let $(\bar{L}_i, \bar{a}_i, \bar{b}_i)$ be the average color of $R^{(i)}$. The following process clusters the point set $\{(\bar{L}_i, \bar{a}_i, \bar{b}_i), i = 1, 2, \ldots, n\}$ and estimates the average color of the board.

1. Define a 3-D window function $\varphi(x,y,z)$ with sides δ_L, δ_a, δ_b and center (x, y, z). $(\bar{L}_i, \bar{a}_i, \bar{b}_i) \in \phi(x, y, z)$ if $|\bar{L}_i - x| < \delta_L$ and $|\bar{a}_i - y| < \delta_a$ and $|\bar{b}_i - z| < \delta_b$.
2. Compute the centroid of the point set:

$$(\bar{L}, \bar{a}, \bar{b}) = \left(\frac{1}{n}\sum_{i=1}^{n}\bar{L}_i, \frac{1}{n}\sum_{i=1}^{n}\bar{a}_i, \frac{1}{n}\sum_{i=1}^{n}\bar{b}_i\right).$$

Initialize the center of the window $\varphi(\cdot)$ as $(\bar{L}, \bar{a}, \bar{b})$.

3. Update the center of the window with the centroid of the points in the window:

$$\bar{L} \leftarrow \frac{\sum_{i,(\bar{L}_i,\bar{a}_i,\bar{b}_i)\in\phi(\bar{L},\bar{a},\bar{b})} \bar{L}_i}{\sum_{i,(\bar{L}_i,\bar{a}_i,\bar{b}_i)\in\phi(\bar{L},\bar{a},\bar{b})} 1},$$

$$\bar{a} \leftarrow \frac{\sum_{i,(\bar{L}_i,\bar{a}_i,\bar{b}_i)\in\phi(\bar{L},\bar{a},\bar{b})} \bar{a}_i}{\sum_{i,(\bar{L}_i,\bar{a}_i,\bar{b}_i)\in\phi(\bar{L},\bar{a},\bar{b})} 1},$$

$$\bar{b} \leftarrow \frac{\sum_{i,(\bar{L}_i,\bar{a}_i,\bar{b}_i)\in\phi(\bar{L},\bar{a},\bar{b})} \bar{b}_i}{\sum_{i,(\bar{L}_i,\bar{a}_i,\bar{b}_i)\in\phi(\bar{L},\bar{a},\bar{b})} 1}.$$

Repeat this process until the center of the window, $(\bar{L}, \bar{a}, \bar{b})$, converges.

The above process is mean-shift in nature. After the computation, $(\bar{L}, \bar{a}, \bar{b})$ is the estimated average color of the pixels in board regions. Although a large amount of statistical techniques for clustering and outlier detection are available, the process described above is sufficient and reliable enough for a coarse estimate of the average board color.

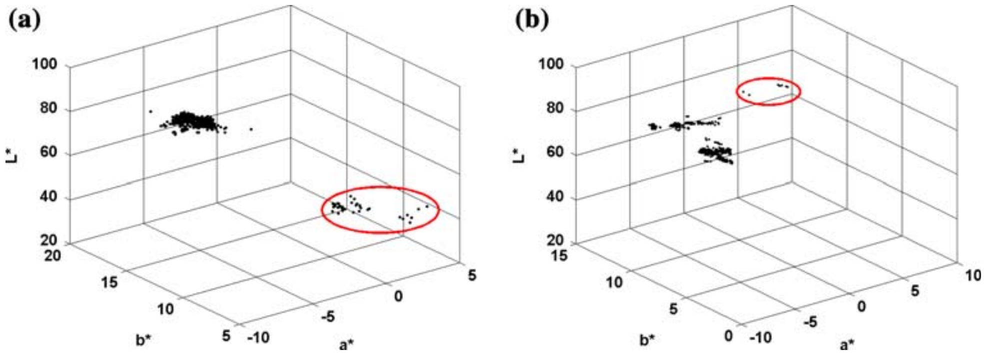

Fig. 2 Distributions of the average colors of the dominant regions in two instructional videos. The points in the ellipses are the outliers—the average colors of the dominant regions that are not content regions. Using our statistical analysis technique, we can separate the outliers and better estimate the variation of color and the range of luminance of the pixels in board regions

The scattering of the largest-region average colors in $L^*a^*b^*$ space is arbitrarily shaped, as shown in Fig. 2. Note that only in a small percentage of frames, the largest regions are not board regions. These outliers are those regions whose average colors are far away from the center of the average colors. The mean shift method is introduced mainly to detect the outliers and better estimate the parameters of the board background.

In the above algorithm, the window dimension (δ_L, δ_a, δ_b) is experimentally selected as (20, 10, 10). After the convergence, the points that are not encompassed by the window are considered as outliers. We tested four instructional videos and found that the window dimensions are not sensitive parameters. As the chalk pixels in the board regions significantly affect the average luminance of the board, we observe that the luminance vary in a range larger than that of the color components. Therefore, δ_L is selected as two times that of δ_a and δ_b.

Based on the average board color $(\bar{L}, \bar{a}, \bar{b})$ and the converged window $\phi(\bar{L}, \bar{a}, \bar{b})$, we further estimate the range of the luminance and the variance of the board color in the $L^*a^*b^*$ color space. Define V as the set of image pixels in the largest regions whose average colors are encompassed by the window, i.e., $V = \cup_{i=1,\ldots,n}\{R^{(i)} | (\bar{L}_i, \bar{a}_i, \bar{b}_i) \in \phi(\bar{L}, \bar{a}, \bar{b})\}$. The standard deviations of the a and b components of the pixels in board regions σ_a and σ_b, are calculated over all the pixels in V, in the straightforward manner. For the L^* component, we compute a cut-off point ℓ such that 95% of the pixels in V have the luminance value in the range [ℓ, $L_{\max}$], i.e.,

$$\frac{\sum_{\mathbf{x} \in V, L(\mathbf{x}) \in [\ell, L_{\max}]} 1}{\sum_{\mathbf{x} \in V} 1} = 0.95,$$

where $L_{\max}$ is the maximum luminance value of all pixels in V and $L(\mathbf{x})$ is the luminance value of pixel $\mathbf{x}$.

The chalk pixels in board regions may have high luminance values. Therefore, we set the upper bound of chalk pixel luminance as the highest possible value, while making a cut-off point at the lower end of luminance. This cut-off point is used to eliminate the outliers. The average board color $(\bar{L}, \bar{a}, \bar{b})$, together with the color distribution parameters ℓ, $L_{\max}$, σ_a, and σ_b, are further used for background separation.

3 Background separation

With the estimated parameters of the board regions, the background separation step accurately separates the board regions from the irrelevant regions. This process is applied to every sampled frame in the video sequence. Suppose a frame is segmented into k regions R_i, i = 1,...,k. For each region R_i, we test whether it is a part of the board region using a probabilistic classifier.

For each pixel $\mathbf{x} \in R_i$, we compute the probability $\Pr(\mathbf{x})$ of the pixel being in the board regions. Let $L(\mathbf{x})$, $a(\mathbf{x})$, and $b(\mathbf{x})$ be the L^*, a^*, and b^* components of the pixel $\mathbf{x}$ in $L^*a^*b^*$ color space. $\Pr(\mathbf{x})$ is defined as

$$\Pr(\mathbf{x}) = \Pr{}^{(L)}(\mathbf{x})\Pr{}^{(a)}(\mathbf{x})\Pr{}^{(b)}(\mathbf{x}),$$

where $\Pr^{(L)}(\mathbf{x}) = 1$ if $L(\mathbf{x}) \in [\ell, L_{\max}]$; otherwise, $\Pr^{(L)}(\mathbf{x}) = 0$. The $\Pr^{(a)}(\mathbf{x})$ and $\Pr^{(b)}(\mathbf{x})$ are modelled as Gaussian probability distributions. The probability $\Pr^{(a)}(\mathbf{x})$ is estimated as

$$\Pr{}^{(a)}(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma_a} e^{-(a(\mathbf{x}) - \bar{a})^2 / 2\sigma_a^2},$$

and $\Pr^{(b)}(\mathbf{x})$ is similarly defined.

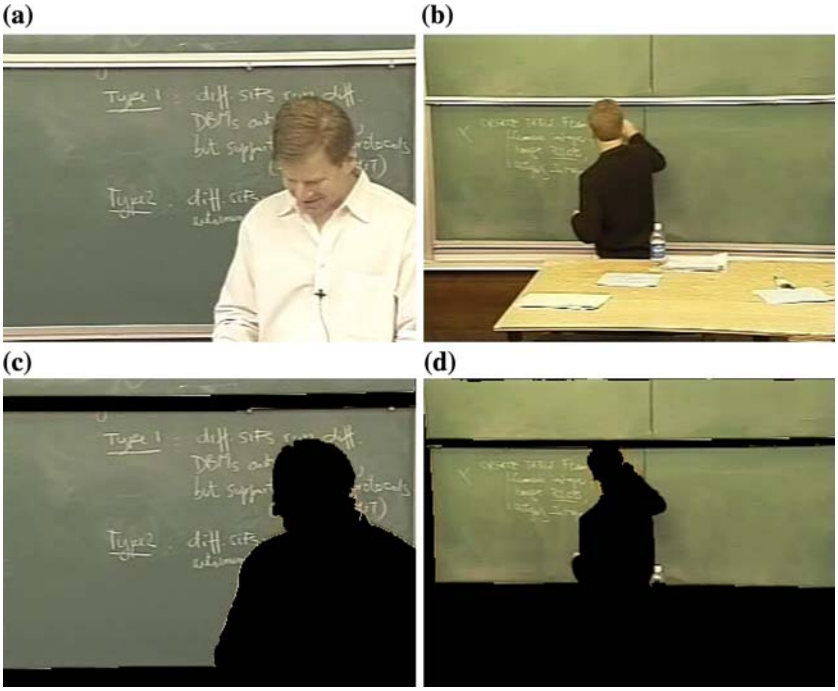

Fig. 3 Sample results of separating content regions from irrelevant regions. **a**, **b** are the original frames; **c**, **d** are the results of background separation. The irrelevant regions (*black*) are separated using our probabilistic classifier and topological refinement

In real instructional videos, we observe that the board pixels show a lot of variation in luminance (L^*) but not much change in their color components (a^* and b^*). Due to the variable amount of chalk dust on the board, the board pixels have luminance values in a large range. In comparison, the color components of the board pixels are distributed in a small range. Therefore, we model the color distribution of the board region as Gaussian distribution.

Then the probability of a region R_i being a content region is defined as:

$$\Pr(R_i) = c \sum_{\mathbf{x} \in R_i} \Pr(\mathbf{x})/|R_i|,$$

where c is a normalization factor that scales $\Pr(R_i)$ to the range of [0, 1]. $c = 2\pi\ \sigma_a\ \sigma_b$. $|R_i|$ measures the total number of pixels in the region R_i. We empirically choose a threshold $\delta_p = 0.3$. If $\Pr(R_i) > \delta_p$, we consider R_i as a content region and merge it with the other content regions; otherwise, we classify it as an irrelevant region. All the content regions are merged together to form the board region:

$$R = \bigcup_{i=1,\ldots,k} \{R_i | \Pr(R_i) > \delta_p\}.$$

The probabilistic model above treats the luminance component (L^*) and color components (a^* and b^*) separately. By allowing a large margin on the luminance component and modelling color components as Gaussian, this model can largely handle the cases of non-uniform board color, light condition change, and the chalk erasure regions. To further improve the accuracy, we make a topological analysis and refine the board region R. If an isolated irrelevant region of less than 512 pixels (in video frames of 320 × 240 pixels) is encompassed by the content regions, we merge it with the surrounding content regions. Similarly, if there exists an isolated content region of less than 512 pixels encompassed by irrelevant regions, we classify it as an irrelevant region. The assumption is that the board regions are not very small, although they may not be topologically connected. This topological analysis can be better understood as a ‘‘hole-filling’’ process. Sample results of the board region separation are shown in Fig. 3..4

4 Content extraction

In instructional videos, the chalk pixels have relatively high luminance values, but they are still close to the board background pixels in luminance and color. Therefore, they cannot be trivially separated by fixed thresholding. Figure 4 shows the 3-D distribution of the pixels of board regions in $L^*a^*b^*$ color space for the frames listed in Fig. 1. These distributions are obtained by manually eliminating the irrelevant pixels in each frame and down-sampling the content region. It is quite clear from Fig. 4 that the transition from chalk

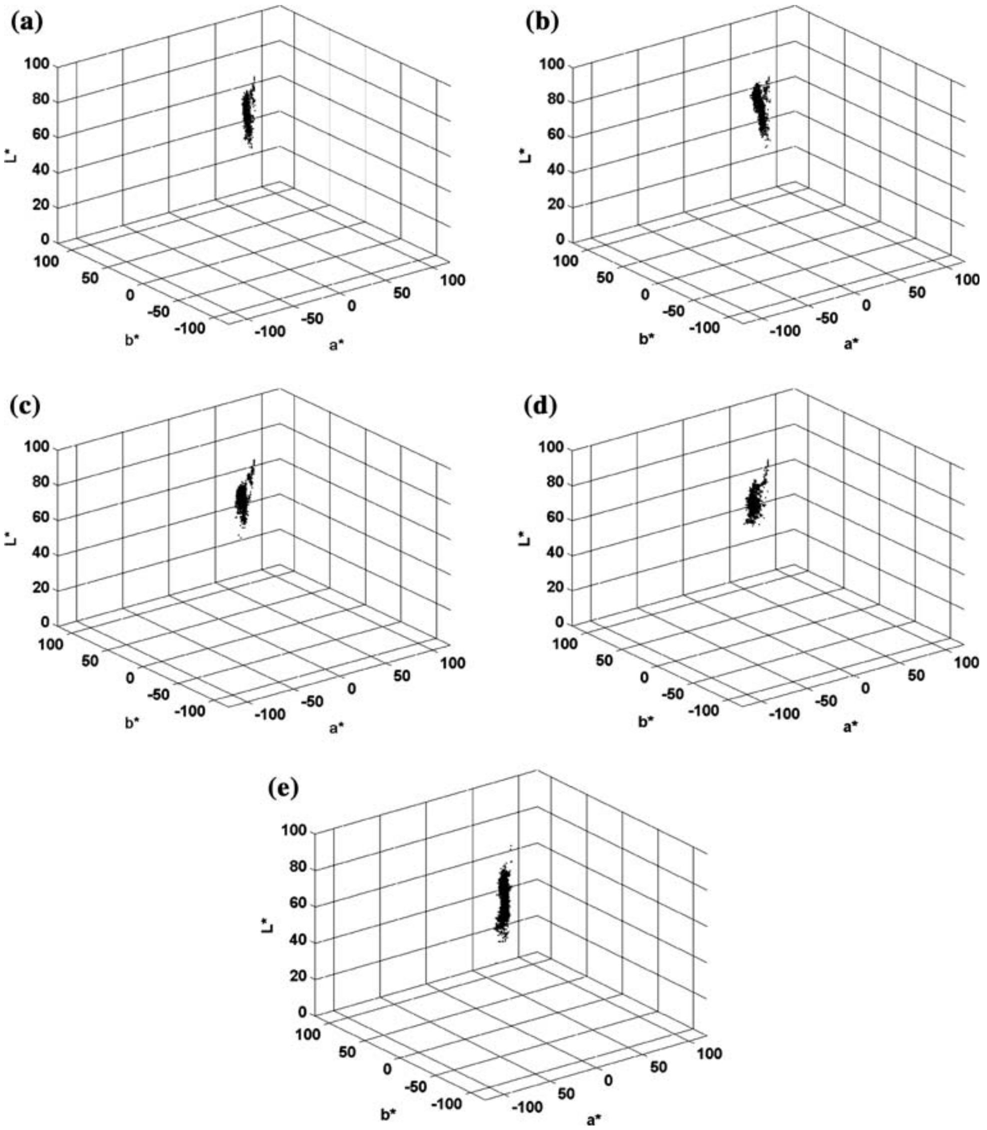

Fig. 4 Distributions of pixel colors in $L^*a^*b^*$ color space. **a–e** are the distributions of the pixels in the board regions of the frames **a–e** in Fig. 1. This figure clearly shows that the pixels in board regions vary significantly in luminance, and a simple thresholding technique cannot separate the chalk pixels from the board background pixels

pixels to board background pixels is gradual and it is impossible to separate them using a fixed threshold. The chalk pixels are ‘‘merged‘‘ with the board background and show more ‘‘flat’’ distributions in luminance compared with the luminance distribution of ink pixels on paper document. So the non-uniformity of the board background needs to be considered in the content extraction process.

The extraction of text and figures in the board regions is based on grey scale morphological filtering followed by gradient-based adaptive thresholding. We first erode the board region, reconstruct the background, and subtract the reconstructed background from the original image by performing the top-hat morphological operation [10]. In our work, we use the Matlab implementation of the top-hat function. The algorithm works as follows. Let R be the image of the board region extracted by background separation. Denote B as a flat structuring element of 10×10 pixels. We erode R using the structuring element B, i.e, $R' = R \ominus B$. If the structuring element is sufficiently larger than the width of the strokes in chalk characters and figures, all the chalk pixels are replaced by the neighboring board background pixels (grey scale erosion is a local minimum operator). For the video frames of size 320×240, we find that a $10 \times\ 10$ flat structuring element effectively removes the chalk content on boards without losing much board background details. Then we iteratively perform geodesic dilations of R' under R until stability is reached. Finally, we subtract the reconstructed board region R' from the original image of the board region: $\tilde{R} = R - R'$. The chalk dust is much larger than the width of the strokes and have relatively low intensity values than the chalk content pixels. The chalk content pixels can be viewed as peaks on the surface plot of the board background. With top-hat processing, the chalk dust is mostly reconstructed and subtracted from the original image, thus enhancing the magnitude of gradient of the chalk pixels.

After the morphological processing, we perform a gradient-based adaptive thresholding on the image $\tilde{R}$. Note that the adaptive thresholding is performed on the morphologically processed image $\tilde{R}$ instead of the original image R. We divide $\tilde{R}$ into blocks of 16×16 pixels. Within each block S, we compute a threshold to extract content pixels. First, we compute the number of edge points in the block:

$$K = \sum_{(x,y)\in S, |\nabla \tilde{R}(x,y)| > \delta} 1,$$

where $|\nabla \tilde{R}(x,y)|$ represents the magnitude of gradient of $\tilde{R}$ derived using the Sobel operator, δ is the threshold for edge point detection. In our work, we use the Matlab function of the Sobel edge detection.

The binarization threshold is selected as the $(1.2K)$-th largest luminance value of the pixels in the block S of the image $\tilde{R}$. We extract all the pixels with luminance values larger than the binarization threshold in that block as content pixels. In addition, a connected component analysis is further applied to remove the connected components that have less than four pixels. This reduces the false-positive content pixels caused by image noise, chalk dust, irregularity of the board, etc.

The binarization threshold is selected with the following consideration. In instructional videos, some chalk pixels have low luminance values and have no strong magnitude of gradient. This happens when instructors write lightly on board or write on the boards whitened by chalk erasures. This can also be understood from the fact that the stroke of a chalk on the boards does not spread the chalk uniformly and leaves some pixels with low luminance values. Therefore, the number of chalk pixels is usually slightly larger than the number of edge points. The factor 1.2 is selected based on our experiments on four instructional videos; it is effective in retrieving most of the chalk pixels.

Note that neither morphological processing nor edge detection alone is sufficient for content pixel detection in instructional videos. The top-hat morphological image processing, which extracts local maxima from images, cannot completely exclude the false content pixels from the chalk erasure regions where pixels tend to have high luminance value. Using edge detection alone cannot handle this case either because the erased chalk pixels still show strong edges locally. With the top-hat morphological image processing, the luminance of the pixels in the chalk erasure regions is greatly attenuated, which leads to weak gradient magnitude that can be handled by gradient-based adaptive thresholding. As shown in Fig. 5, the top-hat morphological processing greatly improves the content extraction result. By combining these two techniques together, we are able to retrieve most of the text and figures while introducing least false-positive pixels.

5 Experimental results

We evaluated the performance of our content extraction algorithm on four full-length instructional videos: (1) video 1, a 77-min video of the course “Computer Architecture”; (2) video 2, a 73-min video of the course “Computer Architecture“; (3) video 3, a 73-min video of the course “Database Systems”; (4) video-4, a 113-min video of the course “Database Systems”. All these videos were recorded in real classrooms by amateur cameramen under poor recording conditions and were used for the distance learning programs in the univer-

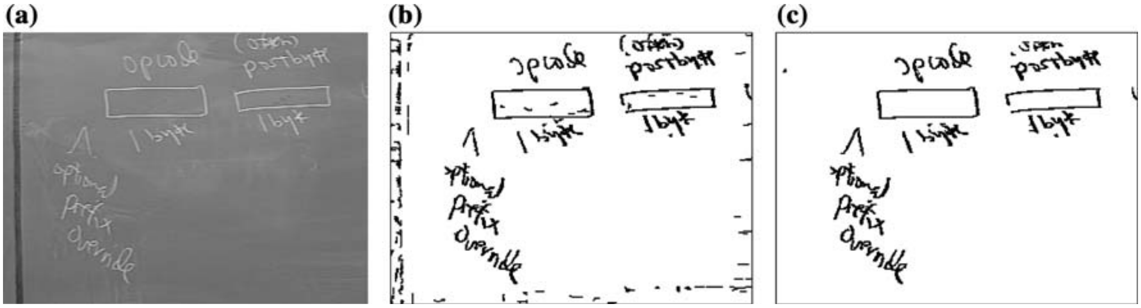

Fig. 5 The top-hat operator attenuates the luminance of the pixels in the chalk erasure regions and reduces the false positives. **a** original grey scale image; **b** content extracted directly by the gradient-based adaptive thresholding (without the preceding top-hat operation); **c** content extracted using the top-hat operator followed by gradient-based adaptive thresholding

Table 1 The performance of content extraction on four full-length instructional videos

Video	Number of sampled frames	False-positive elements < 4			False-positive elements ≥ 4			Time (s)
		$p = 0$	$0 < p < 2\%$	$p \geq 2\%$	$p = 0$	$0 < p < 2\%$	$p \geq 2\%$	
1	884	822	13	7	39	2	1	5701
2	523	481	5	3	32	1	1	3139
3	441	421	6	1	12	0	1	2567
4	1246	1138	6	10	89	0	3	7923

p is the content missing rate. This table shows the number of the frames in each performance range. The last column shows the time taken for processing each video. The average time taken for processing one frame is 6.25 s

sity. They have a significant amount of light condition changes, and occlusions of the board region. In the classrooms one camera was mounted at the back, and camera movements and zooming occurred frequently in the videos. The number of students in the class is 50 on average. Due to the wear of the board and the erasure of written chalk content, the board regions show significant non-uniformity and variation in color and luminance. The instructors used a lot of illustrations, formulae and figures apart from text to teach the course content. In addition, the board has four panels, and the switching of the panels occurred frequently in these videos. All these factors make instructional video content extraction a challenging task.

In all these videos, the instructors mainly use the board to present course content, but the videos also contain other presentation formats (e.g., PowerPoint). Since we focus on board presentation videos in our work, we manually removed the video segments of other presentation forms. In fact, different presentation forms can be easily separated by video sources because they are usually captured by different cameras, and existing methods [20, 27] are available for accurately separating different presentation forms. As instructional videos are highly redundant in video content, we sample these videos at a rate of one frame per 5 seconds (150 frames). Our observation on these four videos indicates no loss of content at this sampling rate.

To evaluate the performance of content extraction, we introduce two concepts: the missing content elements and the false-positive elements. We define a word as a content element for the text; for figures and drawings, we define a stroke (e.g., a line segment) as a content element. Such a content element is the basic unit for high-level applications such as indexing and recognition. With the introduction of the content element, the *content missing rate p* of a frame is defined as:

$$p = \frac{\text{The number of missing content elements}}{\text{The total number of content elements}}.$$

In addition, the extracted content may include non-content (false-positive) pixels. In this case, we define a connected component of irrelevant pixels (i.e., a region of the spatially connected irrelevant pixels) as a *false-positive element*.

Table 1 shows the performance of our algorithm on the four tested instructional videos. The ground truth is obtained by manually looking at the content of each frame and counting the number of content elements. Then, the content extraction result of the frame and the ground truth are compared to count both the extracted and the missing content elements. The content missing rate of the frame is calculated as defined above. We also manually count the number of false-positive elements in each frame. The process of calculating the content missing rate and the number of false-positive elements for a typical frame is illustrated in Fig. 6. For the frame shown in Fig. 6, one content element is missing out of nine. So the content missing rate is 11.1%. Using this methodology, each sampled frame in the video is classified into one of the performance ranges shown in Table 1. About 100 man-hours were spent for evaluating the four listed videos by extracting the ground truth content elements.

In Table 1, we can clearly see that our algorithm achieves high performance on instructional video content extraction. Our algorithm accurately retrieves all content elements in 98% of the video frames of the four tested videos; in 99% of the video frames, the content missing rate is less than 2%. For high-level applications such as indexing and retrieval of instructional videos, a major concern is the loss of the text/figure content [34, 16]. The results in Table 1 clearly show that our algorithm is sufficiently accurate [16] for these applications. Furthermore, our algorithm introduces few false-positive elements. On average, in the four tested videos, the results on more than 94% of the images contain less than four false-positive elements.

The proposed method is also computationally efficient. The system has been implemented in MatLab on a Windows computer equipped with a 2.6 GHz Pentium-4 CPU and 1 GB RAM. The last column in Ta-

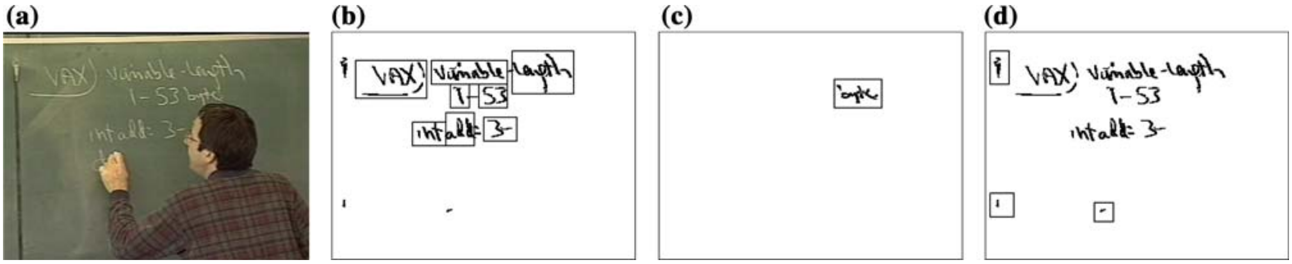

Fig. 6 The content missing rate is measured by comparing the content extraction result with the manually extracted ground truth. In the figure, the content elements and false-positive elements are shown in *rectangular boxes*. **a** The original image; **b** the eight extracted content elements; **c** the one missing content element; **d** the three false-positive elements

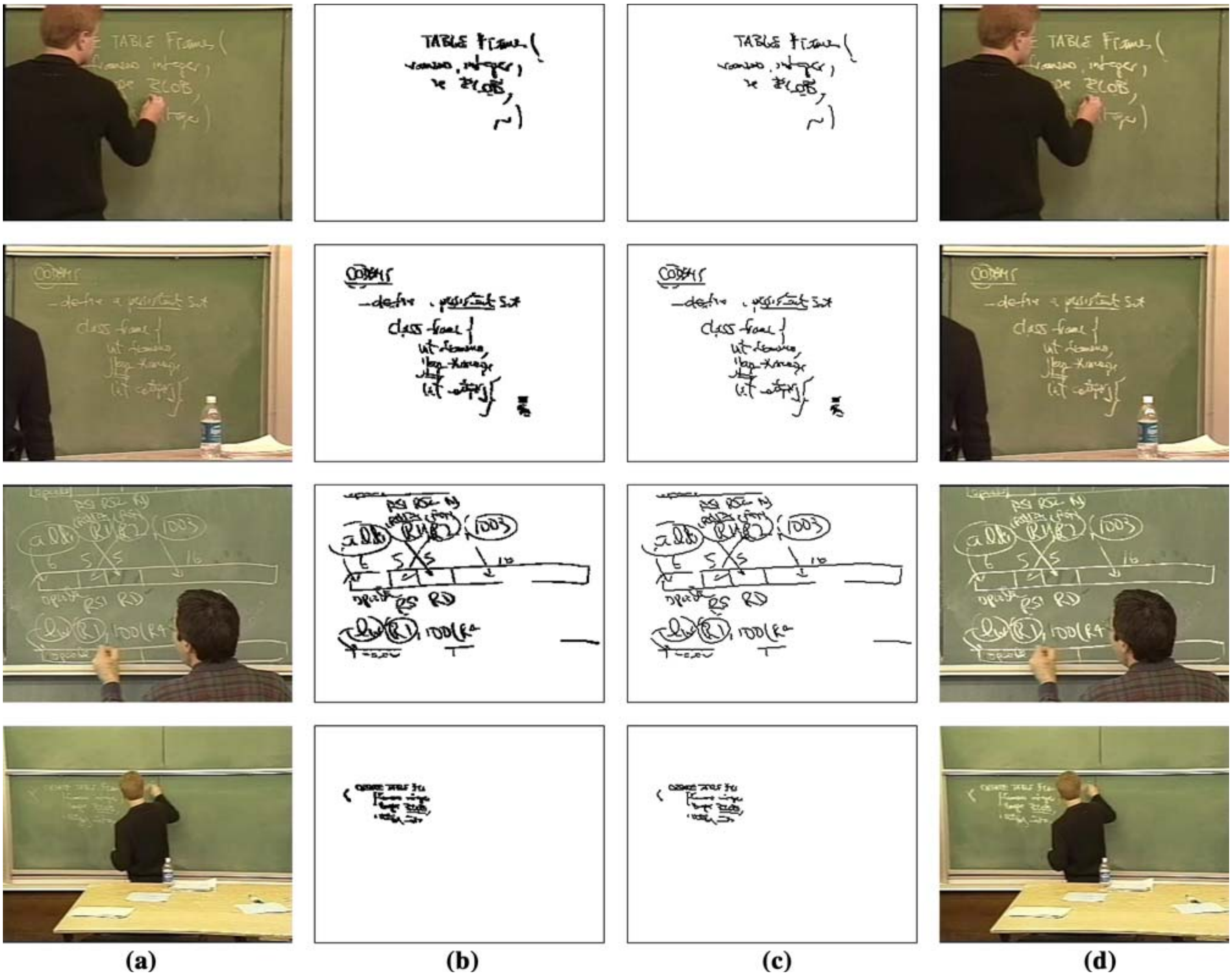

Fig. 7 Sample content extraction results of the instructional videos recorded in real classrooms. *Column a* the original images; *Column b* the extracted content pixels; *Column c* the thinned content text and figures; *Column d* the enhanced video content. Our proposed approach accurately retrieves content text and figures from instructional videos, which enables the development of high-level applications, such as indexing and retrieval of instructional video content

ble 1 shows the time taken for processing each of the four videos. The average time taken for the processing of one frame is 6.25 s.

Many applications can be developed based on our proposed system. For example, the extracted text content can be thinned and fed to handwriting recog-

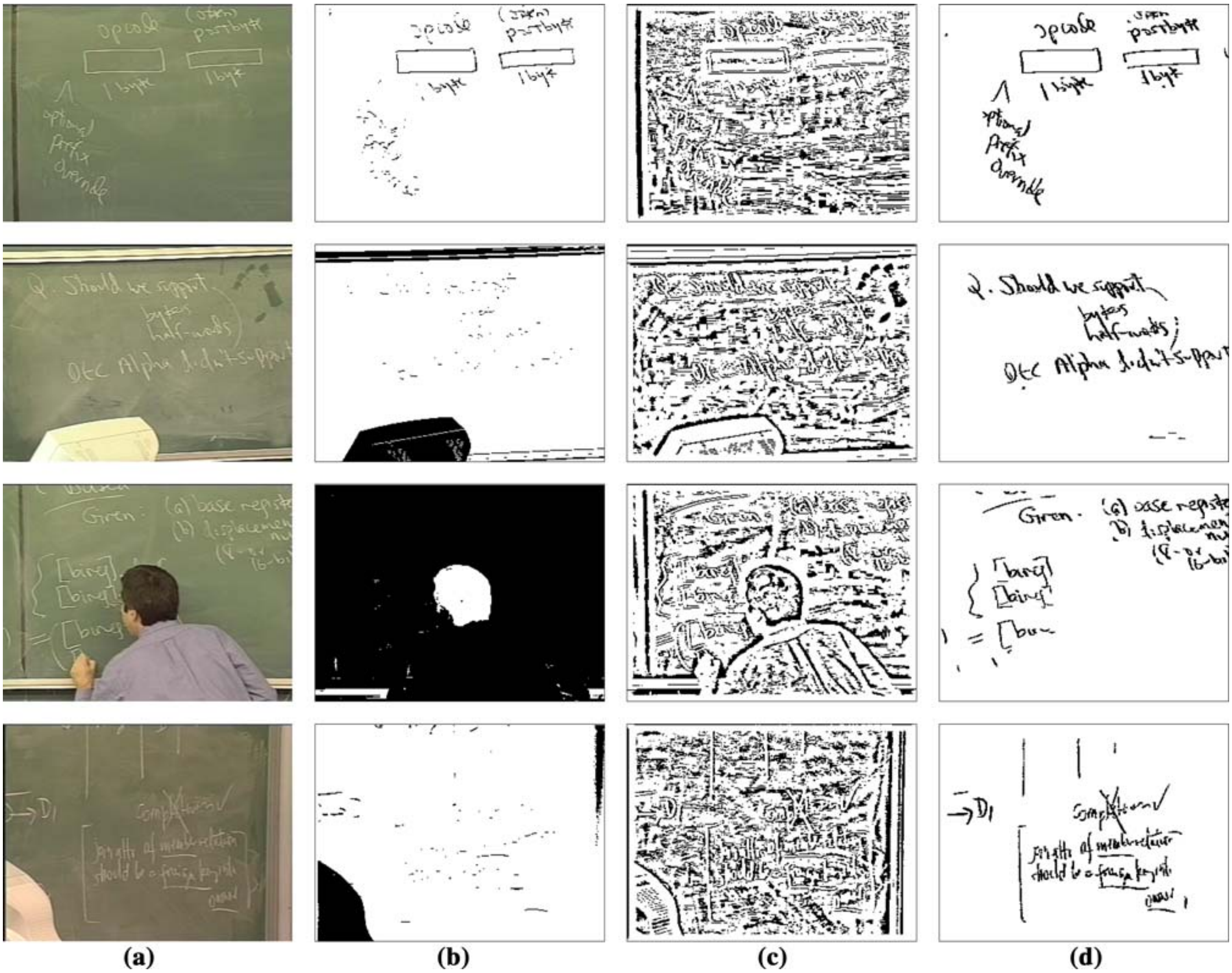

Fig. 8 Comparison of content extraction results by different methods. *Column a* the original images; *column b* the result of Kittler's method [13]; *column c* the result of Niblack's method [25]; *column d* our content extraction result. Kittler's method fails to extract most of the content text, while Niblack's method introduces too many false-positive pixels. In comparison with these two existing methods, our proposed method achieves better performance

nition and information retrieval systems. Figure 7 shows sample images of the experimental results of instructional video content extraction. In Fig. 7c, we show the thinned text and figures, which can be processed by the systems of handwriting recognition and document image analysis. The viewing experience of instructional video content can also be enhanced by stretching the intensity values of the content pixels in board regions, as shown in Fig. 7d.

To show the effectiveness of our algorithm, we compare our results with those of two other techniques in document image processing, namely Kittler's [13] and Niblack's [25] methods. Kittler's method is a minimum-error thresholding technique and is developed for foreground/background separation assuming a single object on the background. In the recent survey paper [28] that evaluates 40 image thresholding techniques and their effectiveness in document image processing, Kittler's method [13] is ranked first in extracting text from document images. Niblack's method is a well known local adaptive thresholding technique. The local threshold in Niblack's method depends on the mean and standard deviation of the neighboring pixels.

Figure 8 shows the comparison of Kittler's method, Niblack's method, and our content extraction method on four sample images. As shown in Fig. 8b, Kittler's method identifies very few content pixels and extracts irrelevant objects in the classroom, like the computer and the bordering frame of the chalkboard. In our algorithm, the background separation step first segments the board background and the succeeding binarization technique extracts most of the content pixels. Niblack's method is very sensitive to the noise on the board background such as chalk dust. Therefore, it introduces a lot of false positives and hardly identifies the content pixels, as shown in Fig. 8c. The text extraction technique presented in this paper is robust to the irregularities of the board background and introduces very few false positives, as shown in Fig. 8d. The content extraction results of different approaches clearly show that our proposed method performs better than these two existing methods.

6 Conclusions and future work

In this paper, we presented a robust method to accurately extract the content text and figures from instructional videos. We analyzed the distribution of content pixels in instructional videos and developed a statistical model for classifying the content pixels. The proposed method first estimates the average and variance of the board color and the range of luminance across the video session. Then for each sampled video frame, it accurately separates the board regions from irrelevant regions by merging image regions using a probabilistic classifier. Finally, it retrieves the content pixels by combining morphological processing with a gradient-based adaptive thresholding technique. The experimental results showed that our algorithm achieved high performance on the four tested full-length instructional videos.

The proposed method can be directly used for enhancing text quality in instructional videos and for handwriting recognition. Furthermore, the content extraction results may aid the analysis of video scenes and the grouping of lecture topics in instructional videos. Future work includes the summarization of instructional videos by combining the extracted content text, the recognition of lecture topics, and the analysis of instructional video scenes.

7 Originality and contribution

Extracting content text from instructional videos of chalkboard presentations is a challenging problem that has not been well explored in the past. In this work, we provide a content analysis method specially developed for chalkboard presentation videos. The contributions of this work are mainly in the following three aspects.

1. This paper provides a content analysis method for instructional videos of chalkboard presentations by adapting the existing image processing techniques to this application. The proposed method is robust to image and video noise, light reflection and non-uniformity of the board, chalk dust caused by erasures, multiple board panels, and occlusions by the instructors.
2. This paper studies and analyzes the statistical distribution of board and chalk pixels. A segmentation refinement technique is introduced to accurately separate the board regions from irrelevant regions in instructional videos.
3. This paper also presents a new binarization technique to extract the text and figures on boards with the consideration of the effect of chalk erasures and the non-uniformity of board regions.

The proposed content analysis method facilitates the research on handwriting recognition, indexing and retrieval, and further analysis of instructional videos.

References

1. Altman E, Chen Y, Low WC (2002) Semantic exploration of lecture videos. In: ACM conference on multimedia, pp 416–417
2. Ankush Mittal SJ, Sumit Gupta, Jain A (2006) Content-based adaptive compression of educational videos using phase correlation techniques. IEEE Trans Multimedia 11(3):249–259
3. Antani S, Crandall D, Kasturi R (2000) Robust extraction of text in video. In: International conference on pattern recognition, pp 831–834
4. Cai M, Song J, Lyu MR (2002) A new approach for video text detection. In: International conference on image processing, pp 117–120
5. Comaniciu D, Meer P (2002) Mean shift: a robust approach towards feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619
6. Davis JL, Smith TW (1994) Computer-assisted distance learning. IEEE Trans Educ 37(2):228–233
7. Dorai C, Oria V, Neelavalli V (2003) Structuralizing educational videos based on presentation content. In: International conference on image processing, vol 3, pp 1029–1032
8. Fan J, Luo H, Elmagarmid AK (2004) Concept-oriented indexing of video databases: toward semantic sensitive retrieval and browsing. IEEE Trans Image Process 13(7):974–992
9. Gao J, Yang J (2001) An adaptive algorithm for text detection from natural scenes. In: International conference on computer vision and pattern recognition, pp 84–89
10. Gonzalez RC, Woods RE (2000) Digital image processing. Addison–Wesley, USA
11. Heng WJ, Tian Q (2002) Content enhancement for e-learning lecture videos using foreground/background separation. In: IEEE workshop on multimedia signal processing, pp 436–439
12. Ju SX, Black MJ, Minneman S, Kimber D (1998) Summarization of videotaped presentations: automatic analysis of motion and gesture. IEEE Trans Circuits Systems Video Technol 8(5):686–696
13. Kittler J, Illingworth J (1986) Minimum error thresholding. Pattern Recognit 19(1):41–47

14. Li H, Doermann D, Kia O (2000) Automatic text detection and tracking in digital video. IEEE Trans Image Process 9(2):147–156
15. Liang J, Doermann D, Li H (2005) Camera-based analysis of text and documents: a survey. Int J Doc Anal Recognit 7(2–3):84–104
16. Lienhart R(1996) Automatic text recognition for video indexing. In: ACM conference on multimedia, pp 11–20
17. Lienhart R, Wernicke A (2002) Localizing and segmenting text in images and videos. IEEE Trans Circuits Syst Video Technol 12(4):256–268
18. Liu T, Hejelsvold R, Kender JR (2002) Analysis and enhancement of videos of electronic slide presentations. In: International conference on multimedia and expo, vol 1, pp 77–80
19. Liu T, Kender JR (2003) Spatial-temporal semantic grouping of instructional video content. In: International conference on content-based image and video retrieval, pp 362–372
20. Liu Y, Kender JR (2003) Fast video segment retrieval by sort-merge feature selection, boundary refinement and lazy evaluation. Comput Vis Image Underst 92(2-3):147–175
21. Malladi R, Sethian JA, Vemuri BC (1995) Shape modeling with front propagation: a level set approach. IEEE Trans Pattern Anal Mach Intell 17(2):158–175
22. Mandal MK, Idris F, Panchanathan S (1999) A Critical evaluation of image and video indexing techniques in the compressed domain. Image Vis Comput 17(7):513–529
23. Mukhopadhyay S, Smith B (1999) Passive capture and structuring of lectures. In: ACM conference on multimedia, pp 477–487
24. Ngo CW, Chan CK (2005) Video text detection and segmentation for optical character recognition. Multimedia Syst 10(3):261–272
25. Niblack W (1986) An introduction to image processing. Prentice-Hall, Englewood Cliffs
26. Onishi M, Izumi M, Fukunaga K (2000) Blackboard segmentation using video image of lecture and its applications. In: International conference on pattern recognition, pp 615–618
27. Phung DQ, Venkatesh S, Dorai C (2002) High level segmentation of instructional videos based on content density. In: ACM confernce on multimedia, pp 295–298
28. Sezgin M, Sankur B (2004) Survey over image thresholding techniques and quantitative performance evaluation. J Electron Imaging 13(1):146–168
29. Stafford-Fraser Q, Robinson P (1996) Brightboard: a video-augmented environment. In: Conference on computer human interface, pp 134–141
30. Syeda-Mahmood T, Srinivasan S (2000) Detecting topical events in digital video. In: ACM conference on multimedia, pp 85–94
31. Tang X, Luo B, Gao X, Pissaloux E, and Zhang H (2002) Video text extraction using temporal feature vectors. In: International conference on multimedia and expo, vol 1, 85–88
32. Wang S, Siskind JM (2003) Image segmentation with ratio cut. IEEE Trans Pattern Anal Mach Intell 25(6):675–690
33. Wienecke M, Fink GA, Sagerer G (2005) Toward automatic video-based whiteboard reading. Int J Doc Anal Recognit 7(2–3):188–200
34. Zhang D, Nunamaker JF (2004) A natural language approach to content-based video indexing and retrieval for interactive e-learning. IEEE Trans Multimedia 6(3):450–458

Author Biographies

Chekuri S. Choudary received his B.Tech. in Electrical Engineering from Jawaharlal Nehru Technological University, India in 2000 and M.E. in Computer Engineering from the University of South Carolina, Columbia in 2002. He is currently a Ph.D. candidate in Computer Science at the Department of Computer Science, University of South Carolina, Columbia. During the year 2002, he worked as a Visiting Research Assistant at University of Southern California's Information Sciences Institute (ISI). His research interests include Image Processing, Multimedia, and E-Learning technologies. He is a student member of IEEE.

Tiecheng Liu received his Ph.D. degree in computer science from Columbia University in 2003. He is an assistant professor in the Department of Computer Science and Engineering at the University of South Carolina. His main research interests include computer vision, image and video processing, multimedia and advanced learning technologies. He has published over 30 refereed papers in the area of computer vision and multimedia technology and served as a committee member for IEEE CBAR'04, CIVR'05, and other conferences. He is a member of IEEE and ACM.