Text Enhancement with Asymmetric Filter for Video OCR

Datong Chen, Kim Shearer and Hervé Bourlard

Dalle Molle Institute for Perceptual Artificial Intelligence
Rue du Simplon 4
1920 Martigny, Switzerland
chen@idiap.ch

Abstract

Stripes are common sub-structures of text characters, and the scale of these stripes varies little within a word. This scale consistency thus provides us with a useful feature for text detection and segmentation. In this paper a new form of filter is derived from the Gabor filter, and it is shown this filter can efficiently estimate the scales of these stripes. The contrast of text in video can then be increased by enhancing the edges of only those stripes found to correspond to a suitable scale. More specifically the algorithm presented here enhances the stripes in three pre-selected scale ranges. The resulting enhancement yields much better performance from the binarization process, which is the step required before character recognition.

1. Introduction

Video OCR aims at integrating of text-based search [1,2] and advanced character recognition technologies, which is now recognized as a key component in the development of advanced video and image annotation and retrieval systems. Text contained in video includes captions and in-vision text (scene text) that are usually names of people, organization, location, date, times and scores, etc.. It is indeed obvious that these text (on top of audio content) provides precise and meaningful information about their content, and consequently useful clues for their indexing and retrieval.

However, text in image and video is far more difficult to detect, localize, and segment than typical text document in clean paper. Video text usually suffers from many types of features making detection and recognition more difficult, including: low resolution, complex background, and noise caused by image acquisition and compression process (e.g., color bleeding). Furthermore, text strings in video can also be multi-colored (e.g., including different color words on the same row), translucent (such as telop text), or with varying shadows and fonts. Conventinal OCR systems that usually require binary images as inputs can not extract text in video frames. Efficient preprocesses, such as text detection and enhancement is necessary to fill the gap between video documents and the input of a standard optical character recognition (OCR) system.

Previous related work on text detection and location in image and video can be classified into region-based methods, texture-based methods and edge-based methods. Region-based methods detect character as the monochrome regions that satisfied certain heuristic constraints. Color or grayscale reducing is necessary to yeilld teh expected uniform text regions using a common image segmentation [11, 9, 10, 4, 8] or color clustering [17]. Texture-based methods locate the text blocks by extracting texture features of text strings. Wu et.al. [5, 6, 7] proposed an algorithm of using statistical properties of the pixel values through out the Gaussian scale space as texture features and classifying the input image pixels into text pixels and background by using a k-means process. Other texture-based methods that employ spatial variance and Haar wavelet as texture features and neural network or simple thresholding as classifiers are presented in [8, 15, 14]. Edge-based methods detect the text by finding vertical edges. In [12], vertical edges are detected by a 3×3 filter and are connected into text clusters by using smoothing filter.

There are both the pixels of text and the pixels of background inside the located text block. An enhancement procedure is necessary to enhance the contrast between text and background so that the text pixels can be segmented easily from the background by using binarization algorithm. In [7], Wu simply smoothes the detected text region to lead a better binarization. Smoothing eliminates noise but can not filter out the background. Therefore, this method can not reliably extract text in complex background, such as video text. In [14, 12], the authors use multi-frame integration to enhance captions in video. The influence of the background is reduced on the basis of motion clues. The multi-frame methods can efficiently enhance the text in video frames with rather different background movements, for example static text with fast moving background, but is not able to clean the background with same or slightly different movements. Sato [12] [13] enhances the text on the basis of its sub-structure: line element, by using filters with four orientations: vertical, horizontal, left diagonal and right diagonal in the located text block. However, because real scales are unknown, it is not possible to design a filter that can enhance the line elements with widely varying widths. All these previous text enhancement work to perform enhancement after the text has been detected and, therefore, can only improve text segmentation.

In this paper, we presents a method for enhancing the text in video using the orientation and scale of local substructure. We locate the sub-structures of the text using edge detection and estimate the orientation and scale of each sub-structure using a family of filters, which is presented in Section 2. We then select three scale ranges and enhance the contrast of these sub-structures as character strokes in each scale range individually to improve the performance of both the text detection and segmentation.

2. Scale Estimation

2.1. Location and selection

Stripes, the common sub-structures of text characters, usually form strong edges against its background in video frames. It can be detected by first finding the edges in the image and then identified by locating its orientation and scale (width). Thus, Canny operator is first employed to detect the strong edges in image (video frame). In general, close points associated with the same edge have similar orientations and scales. Therefore, to reduce the computation, we select only one edge point in a local region to perform the orientation and scale estimation. The image is segmented into $n \times n$ blocks, where n is set equal to half the size of the smallest scale of the substructures of the text (here $n = 2$). The one edge point with the maximum energy is then selected as the candidate point in each block. Commonly, the number of candidate points is about less than 10% of the sum of all pixels in one image. This reduces the number of pixels to be processed in the system, yielding a more efficient algorithm.

2.2. Asymmetric Filter

The points on the edge of a stripe can be identified by estimating the orientations and scales of the possible stripe. As we know, Gabor filters can be designed to be sensitive to stripes of a specified width and orientation. The family of two-dimensional Gabor filters $G_{\lambda,\theta,\varphi}(x,y)$, which was proposed by Daugman [16], are often used to obtain the spatial frequency of the local pattern in an image:

$$\begin{aligned} G_{\lambda,\theta,\varphi}(x,y) &= e^{-\frac{\left(x'^2+\gamma^2 y'^2\right)}{2\sigma^2}} cos\left(2\pi\frac{x'}{\lambda}+\varphi\right) \\ x^{'} &= xcos\theta + ysin\theta \\ y^{'} &= -xsin\theta + ycos\theta \end{aligned}$$

where the arguments x and y represent the pixel coordinaters, parameter θ specifies the orientation of the filter, parameter γ determines the spatial aspect ratio, and $\frac{1}{\lambda}$ is called the spatial frequency.

These filters provide the optimal resolution for both the orientation and the spatial frequency of a local image region. However, the conventional Gabor filter can not extract orientation and scale information of edge pixels since inthis case filter responses are close to zero.

We therefore introduce two groups of Gabor-based asymmetric filters: edge-form filters and stripe-form filters to obtain the precise scale information of the located edges in an image. The edge-form filters $E_{\lambda,\theta}\left(x,y\right)$ are the Gabor filters with $\varphi=\pi/2$:

$$\begin{aligned} E_{\lambda,\theta}\left(x,y\right) &= e^{-\frac{\left(x'^2+\gamma^2 y'^2\right)}{2\sigma^2}} cos\left(2\pi\frac{x'}{\lambda}+\frac{\pi}{2}\right) \\ x^{'} &= xcos\theta + ysin\theta \\ y^{'} &= -xsin\theta + ycos\theta \end{aligned}$$

The stripe-form filters $S_{\lambda,\theta}\left(x,y\right)$ are defined as a Gabor filter with a translation $\left(-\frac{\lambda}{4},0\right)$:

$$\begin{aligned} S_{\lambda,\theta}\left(x,y\right) &= e^{-\frac{\left(x'^2+\gamma^2 y'^2\right)}{2\sigma^2}} cos\left(2\pi\frac{x'}{\lambda}\right) \\ x^{'} &= xcos\theta + ysin\theta - \frac{\lambda}{4} \\ y^{'} &= -xsin\theta + ycos\theta \end{aligned}$$

The rational behind this is that those asymmtric filters can give strong response on candidate edge points in optimal orientation and scale.

The edge-form and stripe-form filters keep most of the properties of the Gabor filters except the specified translation on the position and the phase offset. Figure 1 shows the pattern of the edge-form filters (Fig. 1a) and stripe-form filters (Fig. 1bc) in 8 orientations with $\gamma = 0.92$.

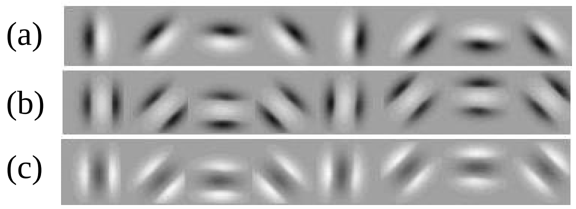


Fig. 1. Asymmetric filters in 8 orientations with $\gamma = 0.92$: (a) edge-form filters , (b) stripe-form filters

These two groups of filters have propertyies that are particularly useful in determining the scale of the local image structure. Experiments show that if the pixel (x,y) is on the edge of stripe structure, the responses of the stripe-form filters are smaller than the responses of the edge-form filters when the scale of the filters are rather smaller than the scale of the stripe ($S_{\lambda,\theta}\left(x,y\right) < E_{\lambda,\theta}\left(x,y\right)$), but greater when the scale of the filters are rather larger than the scale of the stripe ($E_{\lambda,\theta}\left(x,y\right) < S_{\lambda,\theta}\left(x,y\right)$, see Figure 2).

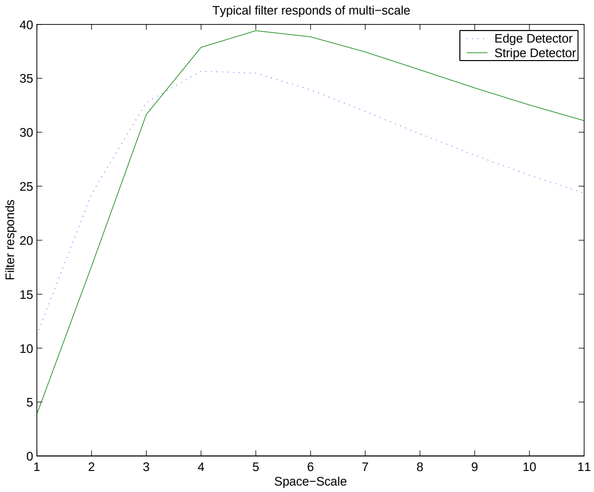


Fig. 2. scale adaptive property of asymmetric filters

The scale at which the stripe-form filter response intersects the edge-form filter response is called the intersection scale, which can be fast located using binary search method. This intersection scale $\overline{\lambda}_s$ roughly indicates the scale of point s. We then train a neural network to model the accurate scale of a candidate edge point s on the basis of five filter responses in scales $\frac{1}{2}\overline{\lambda}_s$, $\overline{\lambda}_s$ and $\frac{3}{2}\overline{\lambda}_s$ in optimalized orientation (the responses of the edge-form filter and the stripe-form filter at intersection scale are the same value). The neural network has only one output α, which indicates the scale λ_s of point s is

$$\lambda_s = \frac{1}{2}\overline{\lambda}_s + \alpha\overline{\lambda}_s.$$

3. Enhancement

These detected scales of candidate edge points may be used to enhance the input video frame before the text is detected so that the enhanced image can benefit both the text detection and segmentation. Since we are not interested with the text in too small or too big size, we select only the candidate edge point, which has scale bigger than 3 and less than 50 meanwhile has enough high responses of stripe-form filters. We then choose three scale ranges to reconstruct the image patterns. The selected candidate points are orgnized into three sets based on their scales. The first set L_1 includes the scales range from 3 to 9, the second set L_2 consists of the scales from 7 to 30 and the last scale level L_3 covers the scales from 26 to 50. Some candidate points may occur in more than one set because of the overlap between L_1, L_2 and L_3. For each candidate point (x, y), with the orientation $\theta_{x,y}$ and filter scale $\lambda_{x,y}$, the three reconstructed image patterns are defined as:

$$I_r^k = \sum_{all\ candidate\ points\ (x,y)} F^k(x, y),\ k = 1, 2, 3.$$

$$F^k(x, y) = \begin{cases} 0 & \lambda_{x,y} \notin L_k \\ S_{\lambda,\theta}(x, y) & \lambda_{x,y} \in L_k \end{cases} .$$

The original image is then enhanced by addition to each of the three reconstructed image patterns I_r^k individually,

$$I_e^k(x, y) = I_{org}(x, y) + I_r^k(x, y),\ k = 1, 2, 3.$$

Figure 3 shows the reconstructed images and the enhanced images. The enhanced image eliminates or blurs the structures that do not have the specified scales while enhances the contrast of the stripes with proper scales. To illustrate the result of this enhancement, we binarized the three enhanced images with Otsu's method [18] and mask off the pixels with zero value in the reconstructed image patterns.

4. Text Recognition

4.1. Text detection

The text detection is based on the edge features in the enhanced image. After detecting edge, we first perform morphological dilation to connect the edges into clusters. For each different scale enhanced image the morphological dilation employs different diameters. The cluster is then bounded into rectangle. If the area of one cluster is smaller than 70% of the area of its rectangle boundary, the cluster is cut into small clusters to ensure that each cluster includes only one dense bar. The candidate text string regions are those bounded clusters which satisfy the constraints that: the horizontal-vertical aspect ration is between 1.2 to 16.0; the height of the cluster is between 6 to 20 for L_1, 15 to 50 for L_2, and 42 to 85 for L_3. Figure 4 shows a sample of text detection.

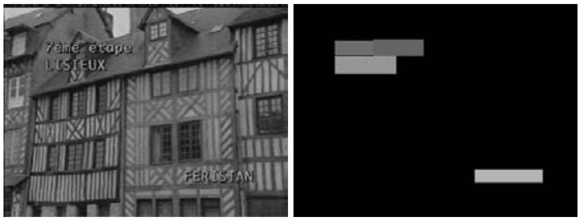


Fig. 4. original image and detected regions in L_2

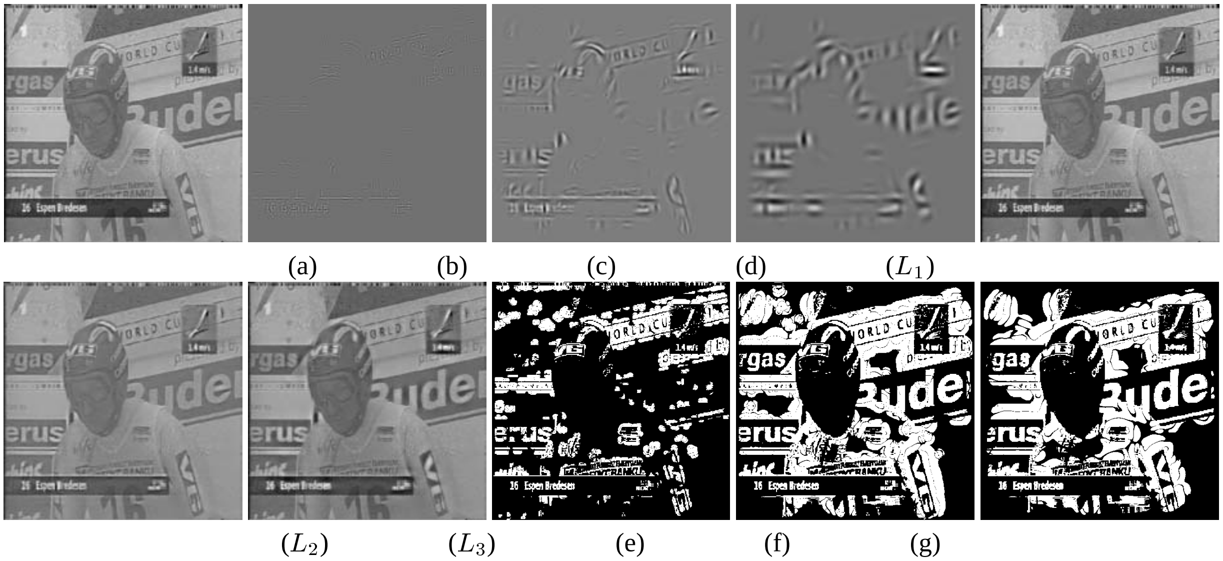
4.2. Text recognition

It is likely that using explicit segmentation then an OCR system designed for documents does not provide the optimum solution for image and video text recognition, however, they do provide a readily available solution. In addition to this, document OCR system are moving rapidly to embrace image text recognition, so while an approach without explicit segmentation may eventually provide better results, standard OCR systems should remain the most useful solution for sometime. In order to use a standard OCR system, the detected text region in enhanced image is normalized to the same height of 128 pixels with bilinear interpolation and then directly binarized to segment the text and background. Figure 5 shows the results of binarization of the text region in original image and enhanced image. A commercial OCR package is then employed for final character recognition.

Table 1. Recognition results: Text 1: superimposed text. Text 2: scene text

algorithm	text	frames	recognition rate
original	1	4000	36.1%
original	2	4000	7.4%
enhanced	1	4000	82.6%
enhanced	2	4000	13.4%

Experiments are based on 4000 frames with different sizes of text in each frame. The text can be superimposed text or scene text, which involves total of 52112 characters. The final recognition is performed by using TypeReader OCR package [19]. We tested the text detection and segmentation on both original frame images and en-



(a) (b) (c) (d) (L_1)

(L_2) (L_3) (e) (f) (g)

Fig. 3. (a) is original frame image; (b, c, d) are reconstructed image patterns in 3 scale ranges; (L_1, L_2, L_3) are enhanced images in 3 scale ranges; (e, f, g) are the binarization results of the 3 enhanced images, using Otsu's method.

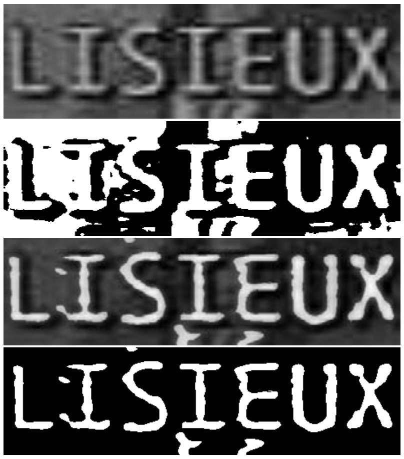


Fig. 5. (a) original image. (b) binary image. (c) enhanced image. (d) binary enhanced image

hanced images using the method presented in this paper. Final recognition results are reported in Table 1. The proposed enhancement yields better recognition performance for both these two types of text. The low recognition rate for scene text results from the fact that the scene text has different alignments, which is then more difficult to detect.

5. Conclusion

Two groups of asymmetric Gabor filters have been proposed which can efficiently extract the orientation and scale of the stripes present in a video image. This information is used to enhance contrast at only those edges most likely to represent text in the scale interest of. The experimental results show that the approach presented in this paper can improve the recognition rate of superimposed text significantly.

6. References

[1] M. Bokser, "Omnidocument technologies", Proc. IEEE, 80(7):1066–1078, July 1992.

[2] S. V. Rice, F. R. Jenkins, and T. A. Nartker. "OCR accuracy: UNLV's fifth annual test", INFORM, 10(8), September 1996.

[3] L. O'Gorman and R. Kasturi, "Document Image Analysis", IEEE Computer Society Press, Los Alamitos, 1995.

[4] J. Ohya, A. Shio, and S. Aksmatsu, "Recognition characters in scene images. IEEE Trans. Pattern Analysis and Machine Intelligence", 16(2):214–220, 1994.

[5] V. Wu, R. Manmatha, and E. M. Riseman, "Finding text in images", In Proc. ACM Int. Conf. Digital Libraries, 1997.

[6] V. Wu, R. Manmatha, and E. M. Riseman, "Textfinder: An automatic system to detect and recognize text in images", IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1224–1229, 1999.

[7] V. Wu and R. Manmatha, "Document image clean-up and binarization", In Proc. SPIE Symposium on Electronic Imaging, 1998.

[8] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images", Pattern Recognition, 28(10):1523–1536, 1995.

[9] R. Lienhart and W. Effelsberg, "Automatic text segmentation and text recognition for video indexing", Technical Report TR-98-009, University of Mannheim, Mannheim, 1998.

[10] R. Lienhart, "Automatic text recognition in digital videos", In Proc. SPIE, Image and Video Processing IV, January 1996.

[11] R. Lienhart, "Indexing and retrieval of digital video sequences based on automatic text recognition", In Proc. 4th ACM International Multimedia Conference, Boston, November 1996.

[12] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video ocr for digital news archives", In IEEE Workshop on Content Based Access of Image and Video Databases, Bombay, January 1998.

[13] T. Sato, T. Kanade, E. K. Hughes, M. A. Smith, and S. Satoh, "Video OCR: indexing digital news libraries by recognition of superimposed caption", In ACM Multimedia System Special Issue on Video Libraries, Feb. 1998.

[14] H. Li and D. Doermann, "Text enhancement in digital video using multiple frame integration", ACM Multimedia 1999.

[15] A. K. Jain and B. Yu, "Automatic text localization in images and video frames", Pattern Recognition, 31(12):2055–2076, 1998.

[16] G. Daugman, "Uncertainty relations for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters", Journal of the Optical Society of America A, 1985, vol.2, pp. 1160-1169.

[17] K. Sobottka, H. Bunke, H. Kronenberg, "Identification of text on colored book and journal covers", ICDAR, pp: 57-63, 1999.

[18] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, 9(1), pp: 62-66, 1979.

[19] TypeReader, www.expervision.com