IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse

Multiple Hypotheses Video OCR

Datong Chen, Juergen Luettin
IDIAP, Switzerland
{chen, luettin}@idiap.ch

IDIAP–RR 00-28

July 2000

PARU DANS
DAS'2000

Institut Dalle Molle
d'Intelligence Artificielle
Perceptive • CP 592 •
Martigny • Valais • Suisse

téléphone +41−27−721 77 11
télécopieur +41−27−721 77 12
adr.él. secretariat@idiap.ch
internet http://www.idiap.ch

Rapport de recherche de l'IDIAP 00-28

Multiple Hypotheses Video OCR

Datong Chen, Juergen Luettin
IDIAP, Switzerland
{chen, luettin}@idiap.ch

July 2000

paru dans
DAS'2000

Résumé. In this paper, we present a method to improve video OCR with multiple character hypotheses. The text regions in video need to be binarized before work as the input of current OCR system. Tranditional binarization do not use any structural information about the text. Based on a certain statistic model, we define a binarization method, which is called observation function, that should satisfy a certain condition. We then present a method to construct an observation function by computing binarization results according to multiple hypotheses of characters obtained by an OCR system..

1 Introduction

Combining video OCR system and text-based search engine for video indexing and retrieving has become an important research topic in recent years. Text information contained in video provides precise and explicit meanings. Superimposed captions usually provide information about the name of people, organization, location, subject, date, time and scores, etc. These texts are powerful resources for indexing, annotation and content-oriented video coding and processing. Moreover, text-based searching and matching technologies have well developed. Text-based search engines address the problem of automatically and efficiently finding documents. Many efficient search engines are employed in peer visiting web sites, e. g. Yahoo and Alta Vista, libraries, and so on.

The OCR systems for machine printed text on clean papers yield high recognition rates and are now readily available for personal computers [1][2]. However, the existing OCR technologies are not able to recognize text that is not printed against a clean background. In videos, the background can be any kind of indoor or outdoor scene while the text can vary in font, size, color, texture, alignment, 3D position and movement, lighting condition, and shading. Therefore, there is high demand for systems to detect unconstrained text from any background, which is called "video OCR system".

Former work on text detection in video can be classified as region-based methods and texture-based methods. In region-based methods, first, the input video frames are roughly segmented into regions. Then the text regions are selected by checking the features, such as maximum and minimum of the region, the width/height ratio of the region, contrast between the segments and the background, spatial frequency of the region, as well as certain alignment and movement. [4][11][9][10] The region-based text detection methods have high processing speed and localization accuracy, however, require the gray-level of the text to be similar.

Texture-based methods regard the text as a spacial kind of texture and employ texture analysis to search the text in the video. In [5][6][7] is an algorithm by using statistical properties through out the Gaussian scale space. Under this scheme, the input image is segmented into text regions and background. The text regions are refined under heuristic constraints to text strings, such as height similarity, spacing and linear alignment, fixed spatial frequency. In [8], spatial variance of the input image is used as feature for locating the text. In [14], the Haar wavelet transform is employed to give out the texture feature. The texture-based methods are robust to small gray-level variation of the text while generally time-consuming and sensitive to character font size and style. Another disadvantage of texture-based methods is that it can not always locate the text accurately[15].

Both the region based and texture based methods try to locate the text regions at first and binarize these regions afterwards to satisfy the input requirement of the current OCR system. In this paper, we present an approach to improve the performance of the video OCR, which offers multiple hypotheses instead of one binarization result as the inputs of the OCR system.

2 Statistical model

The text in video usually has different kinds of local background, which presents the main problem in video OCR. This problem can be described formally with a statistical model. Let I_0 indicate the set of all possible input images of video frames, $|I_0|$ represents the number of images in I_0. We assume that the resolution of all input images or video frames is uniform. C_0 is the set of all possible characters and their positions in image. If the occurrence of a character $c \in C_0$ in image $I \in I_0$ is distinguishable, we say that c appears in I and write $c \in I$. We then define the occurrence probability of c as $p(c)$.

$p(c) = \frac{|I|}{|I_0|}, I = \{I_k \,|I_k \in I_0, c \in I_k\}$

The video OCR is to find C^*

$$C^* = \arg\max_C \left(p\left(I|C\right)\right)$$

with

$$p\left(I|C\right) = \begin{cases} 0 & C \notin I \\ \frac{1}{|I|} & C \in I, I = \{I_k \,|I_k \in I_0, C \in I_k\} \end{cases}$$

The output of an existing OCR system may offer us the probability $p\left(C|\theta\left(I\right)\right)$, here $\theta\left(I\right)$ is the observation function of an OCR system on the image I. Obviously, if for

$\forall C_1, C_2 \in 2^{C_0}$,

$\frac{p(C_1|\theta(I))}{p(C_1)} < \frac{p(C_2|\theta(I))}{p(C_2)}$

$\Rightarrow \frac{p(C_1|I)}{p(C_1)} < \frac{p(C_2|I)}{p(C_2)}$ (1)

we have

$$\arg\max_{C} \frac{p(C|\theta\,(I))}{T \cdot p(C)} \Rightarrow \arg\max_{C} \left(p\,(I|C)\right) \qquad (2)$$

where the coefficient T is a constant. Function θ is called observation function, which satisfies the condition specified in equation (1). We have now a joint between current OCR and video OCR. This formulation combines both the image evidence and the observation function at the OCR system to segment and recognise text.

3 Design of the Observation Function

Tranditional binarization process can be regarded as an observation function if it can extract most of the text pixels from the background. Unfortunately, previous work has shown that it is not possible to segment the pixels of the text without knowing where and what the characters are [11][15]. To address this problem, we preprocess the text region with multiple hypotheses of characters and positions before binarization. The idea is to smooth the different parts of the given region according to the standard character patters, so that we can keep the edge of the character while smooth the inner of the characters. This preprocess can be described as: for each character $c \in C_0$, we design a 2-D weight matrix W_c. If we predict a character $c \in C_0$ in the region X, we first resample the region X to the same size of W_c, so that each pixel in region X corresponds to one element in W_c. Then, for each pixel in region X, we compute its new value by convoluting the region X with a kernel specified by its correspond in weight in W_c. For each pixel $x_i \in X$, the kernel $\eta(x_i)$ is defined as:

$$\begin{aligned}\eta\,(x_i) &= \frac{1}{W_c(x_i)} \exp\left(-\frac{\pi^2 \cdot (x - x_i)^2}{2 \cdot W_c^2(x_i)}\right) \\ &\quad \cdot \cos\left(-\frac{\alpha \cdot \pi \cdot (x - x_i)^2}{W_c^2(x_i)}\right)\end{aligned}$$

where α is a constant parameter. We denote the old value of the pixel x_i as $I(x_i)$, and the new value of the pixel x_i as $I^{'}(x_i)$.

$$I^{'}\,(x_i) = \frac{\sum_{x \in X} I(x_i)\eta\,(x_i)}{\sum_{x \in X} \eta\,(x_i)}$$

W_c can be adjusted manually according to the training data so that we can let the final binarization result roughly satisfy the condition (1).

4 Experiment and Result

We performed an experiment based on the video streams with totally 4000 frames from 3 videos. The texts in the video are both superimposed text and scene text with different alignments and movements. We employed a weight matrix for three kinds of fonts of English characters. The candidate text regions are extracted using a region based method. Here, we constrain that the minimum size of the region should be 8 pixels and this region should be available in at least 5 continuous frames. Figure 1 shows the results of our algorithm and the results with using only a binarization method.

5 Conclusion

The results show that the multiple hypotheses approach presented in this paper can improve the recognition rate of the superimposed text and reduce the error rate, which denotes the error rate of the OCR system. The performance of the preprocess, in which is presented in section 3, is sensitive to the transforms and fonts of the character. This may cause the low recognition rate for the scene text. Because we have not predicted the possible characters but only the position of the characters, the algorithm is rather time-consuming. Future work will focus on the character prediction.

Références

[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.

TAB. 1 – *Recognition results*

Algorithm	type of text	frames	recognition rate	error recognition rate
directly binarization	superimposed text	4000	36.1%	84.7%
directly binarization	scene text	4000	7.4%	—-
multiple hypotheses	superimposed text	4000	70.6%	27.0%
multiple hypotheses	scene text	4000	9.2%	—-

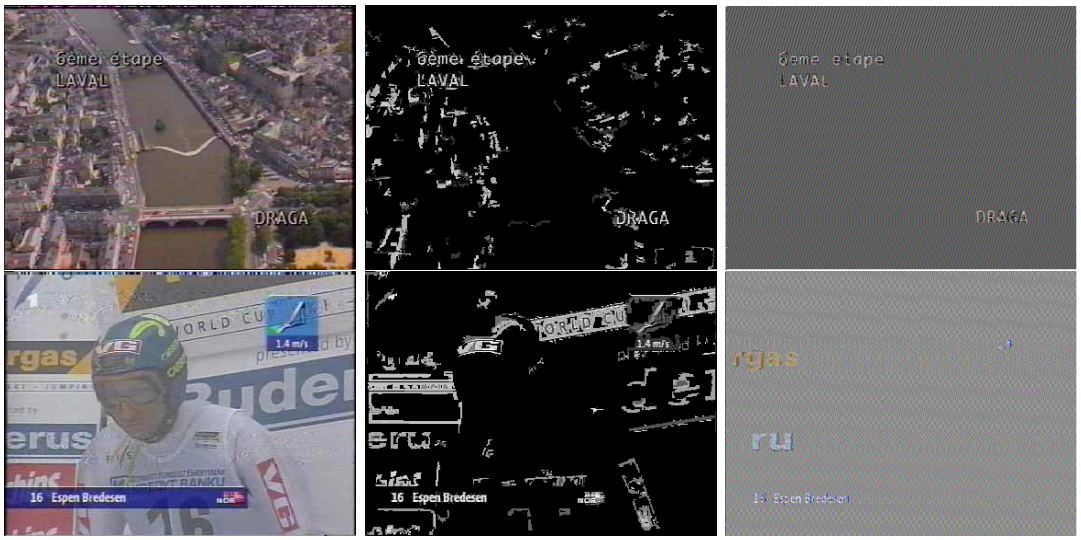

FIG. 1 – *experiment results: (left) original frame image; (middle) candidate regions; (right) observation result*

[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 localisation in images and video frames", Pattern Recognition, 31(12):2055–2076, 1998.