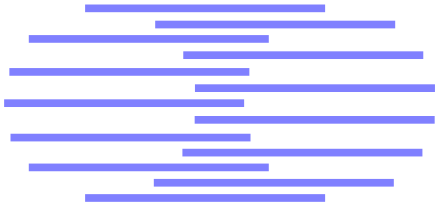
IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse



Text Recognition in Complex Background Based on Markov Random Field

Datong Chen and Jean-Marc Odobez
IDIAP, Switzerland

IDIAP–RR 01-47

Dec. 2001

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 01-47

Text Recognition in Complex Background Based on Markov Random Field

Datong Chen and Jean-Marc Odobez
IDIAP, Switzerland

Dec. 2001

Résumé. In this paper we propose a method to segment and recognize text embedded in video and images. An Expectation-Minimization (EM) procedure with mixture of gaussian is first employed for segmentation. Then, prior in the segmentation is introduced, through a Markov Random Field (MRF) model whose parameters are estimated on-line and therefore adapted to the specific text image. Additionally, text hypothesis are made from each gaussian layer, before being processed by an optical character recognition (OCR) software. Final result is selected from the OCR output, allowing us to deal with text of any grayscale value. Merits of the proposed scheme is assessed from one hour of sports video, and show that the introduction of the prior and of the hypothesis generation improve the performances.

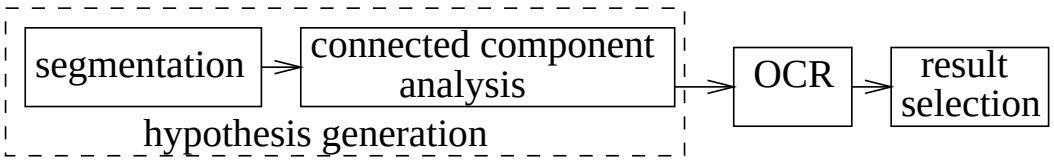


FIG. 1 – *Text recognition scheme*

1 Introduction

Text recognition in video and images aims at integrating text-based search and advanced OCR technologies. It is now recognized as a key component in the development of advanced video and image annotation and retrieval systems. Text characters contained in video are of low resolution, of any grayscale value (not always white) and embedded in complex background even when the whole text string is well located. Thus, applying conventional OCR technology directly leads to poor recognition rate. Efficient segmentation of text characters from background is therefore necessary to fill the gap between video documents and the input of a standard OCR system.

Previous work on text segmentation from complex background have been published in recent years. Lienhart [9] and Bunke [11] clustered text pixels from images using a common image segmentation or color clustering algorithm. Although these methods can somehow avoid text location work, they are very sensitive to noise and character size. Most top down text segmentation methods are performed after text string is located in images. These methods assume that the grayscale distribution is bimodal and that characters correspond a priori to either the white part or the black one, but without providing a way of choosing the right one on-line. Great efforts are thus devoted to perform better binarization, combining global and local thresholding [1], or M-estimation [7], or simple smoothing [12]. However, these methods are unable to filter out background regions with similar grayscale value to characters. Text enhancement methods, if the character grayscale value is known, can help the binarization process [10]. However, without estimation of scales, the designed filters can not enhance character stroke with varying width [4]. In video, multi-frame enhancement [8] also can reduce the number of background regions, but only when text and background have different movements.

In this paper, we present a method based on multi-hypothesis generation and MRF to improve segmentation and recognition of embedded text with unknown grayscale value. Indeed, in Section 3, we compare the results of different hypothesis generation algorithms that are presented in the following Section.

2 Description of the method

We located text using former work presented in [3]. In this method, text-like textures are first detected by integrating horizontal and vertical edges, and further segmented into single line text candidates using baseline location. A support vector machine is then used to identify text regions from the candidates.

This text location step provides us with text images as those presented in figure 2. As we mentioned before, OCR software can not be applied directly. Indeed, experience shows that OCR performances are quite unstable, as already mentioned by others [9], and significantly rely on the segmentation quality, in the sense that errors made in the segmentation are directly forwarded to the OCR. In our case, we propose a softer scheme (see figure 1) in which multiple text layer candidates are provided to the OCR, delaying the hard decision, if any, after the OCR step. Our algorithm works as follows: first, text hypothesis are generated, relying on a segmentation step followed by a connected component analysis; then hypothesis are processed by the OCR and the result is selected from the output.

2.1 Segmentation methods

Let S denote the set of sites s (pixels), and o the observation field $o = \{o_s, s \in S\}$, where o_s corresponds to the gray-level value at site s. We model the image intensities in terms of the combination of K simple random processes, also referred to as layers. Each simple process is expected to represent regions of the image having similar gray levels, one of them being text. Thus, the segmentation consists in the mapping of pixels to processes. It is stated as a statistical labeling problem, where the goal is to find the label field $e = \{e_s, 1 \leq e_s \leq K, s \in S\}$ that best accounts for the observations, according to a given criterion.
To perform the segmentation, we tested 3 algorithms. In all 3 cases, the probability that a gray value arise at a given site, within a particular layer i, is modeled by a gaussian, i.e. $\mathrm{p}_i(o_s) = \mathcal{N}(m_i, \sigma_i)$.

2.1.1 The basic EM algorithm

Here, individual processes are combined into a probabilistic mixture model according to:

$$\mathrm{p}(o_s) = \sum_{k=1}^{K} \mathrm{p}(o_s | e_s = k)\mathrm{p}(e_s = k) = \sum_{k=1}^{K} \pi_k \mathrm{p}_k(o_s) \tag{1}$$

Given an image, the goal is thus to find the set of parameters $\varphi = (\mu_i, \sigma_i, \pi_i)$ which provides a maximum likelihood fit to the data set o, i.e. which maximizes $\mathrm{L}^{\varphi}(o) = \sum_{s \in S} \ln \mathrm{p}(o_s)$. To do so, we can use the standard EM algorithm [5], which consists in iteratively maximizing the expected log-likelihood of the complete data (o,e) with respect to the unknown data (e). After maximization, labels are assigned to the individual layer according to:

$$\forall s \in S \qquad e_s = \underset{i}{\operatorname{argmax}} \ \mathrm{p}_i(o_s) \tag{2}$$

2.1.2 Markov random field assignment

While the EM algorithm is able to capture most of the gray level distribution properties, it does not model the spatial correlation between assignment of pixels to layers, resulting in noisy label images. To overcome this, we introduce some prior, modeling the label field as a MRF. Then, instead of using the simple rule (2), we perform a Maximum a-posteriori (MAP) optimization. Due to the equivalence between MRF and Gibbs distribution ($\mathrm{p}(e) = \frac{1}{Z(V)} e^{-U_1^V(e)}$) [6], this is equivalent to the minimization[1] of an energy function $U(e,o) = U_1^V(e) + U_2^{\varphi}(e,o)$ with $U_2^{\varphi}(e,o) = \sum_{s \in S}(-\ln \mathrm{p}_{e_s}(o_s))$ expressing the adequacy between observations and labels, as in the EM algorithm, and U_1 is equal to:

$$\sum_{s \in S} V_{11}(e_s) + \sum_{<s,t> \in \mathcal{C}_{hv}} V_{12}^{hv}(e_s, e_t) + \sum_{<s,t> \in \mathcal{C}_{diag}} V_{12}^{d}(e_s, e_t) \tag{3}$$

where $\mathcal{C}_{hv}$(resp. $\mathcal{C}_{diag}$) denotes the set of two neighbor pixels in the horizontal, vertical (resp. diagonal) direction. The V are the (local) interaction potentials which expresses the prior on the labels. For instance, assuming layers to be spatially homogeneous, we can select the potentials according to ($V_{11}(i) = 0$):

$$V_{12}^{hv}(k,j) = \beta_{hv}(1 - \delta_{k=j}), V_{12}^{d}(k,j) = \beta_d(1 - \delta_{k=j}) \tag{4}$$

where δ denotes the kronecker function. β_d and β_{hv} are therefore costs to pay to have different labels at neighboring pixels in the corresponding direction.
To summarize, in this method, after the EM algorithm, the label assignment is made by minimizing the energy $U(e,o)$ with a fast ICM procedure [6], where the regularisation parameters V are set to a fixed value.

1. with respect to e

2.1.3 The Gibbsian EM algorithm (GEM)

One may wish to learn the parameters off-line, from examples. However, this would require to know the correspondence between learned labels/layers and current ones.[2] The 3rd algorithm we propose consists in estimating all the parameters Θ=(μ_i,σ_i,V) using an EM procedure. Recall that the E step involves the computation of:

$$\mathbb{E}\left[\ln \mathrm{p}_{eo}^{\Theta} | o,\Theta^n\right] = \sum_e \ln\left(\mathrm{p}_{o|e}^{\Theta}(e,o)\mathrm{p}_e(e)\right)\ \mathrm{p}_{e|o}^{\Theta^n}(e,o) \tag{5}$$

which is then maximized over Θ. Two problems arise here. Firstly, this expectation on $\mathrm{p}_{e|o}^{\Theta^n}$ can not be computed explicitly nor directly. Instead, this law will be sampled using Monte Carlo methods with a Gibbs sampler, and the expectation will be approximated along the obtained Markov chain. Secondly, the joint log-likelihood probability p_{eo}^{Θ} is not completely known, because of the presence of the uncomputable normalizing constant $Z(V)$ in the expression of $\mathrm{p}(e)$. To avoid this difficulty, we use the pseudo-Likelihood function [2] as a new criterion, that is in [6], we replace $\mathrm{p}(e)$ by its pseudo-likelihood $\mathrm{p_S}(e)$ defined from the conditional probabilities:

$$\mathrm{p_S}^V(e) \doteq \prod_{s\in S} \mathrm{p}(e_s|e_{\mathcal{G}_s}) \tag{6}$$

where $e_{\mathcal{G}s}$ represents the label in neighborhood of s. Using this new criterion, the maximization of the expectation (5) can be executed, providing new estimates of (μ_i,σ_i) and of V. The reader is referred to [2] for more details on the procedure that we have adapted to our need. Complexity of the GEM algorithm is approximately 3 times of EM complexity, with non optimized code.

2.1.4 The initial hypothesis generation

After segmentation, for each label, a binary text image hypothesis is generated by assuming that this label corresponds to text and all other labels corresponds to background. Therefore, if we have K different labels, we get K hypothesis. The right choice of K is another important and difficult issue. One general way to address the problem consists in checking whether the increase in model complexity really provides a better fit of the data. This can be done for instance by using the minimizing description length criterion. However, this information theoretic approach may not be appropriate for qualifying a good text segmentation. Therefore, we use a more conservative approach, by varying K from 2 to 4, generating in this way nine text image hypothesis.

2.2 Connected component analysis

A simple connected component analysis is used to eliminate non character regions in each hypothesis to help the OCR system. We keep only connected component that satisfies the following constraints: size is bigger than 5 pixels; width/height ratio is between 4.5 and 0.1; the width of the connected component is less than 2.1 of the height of the whole text region.

2.3 Result selection and confidence value

Each text candidate is then processed by the OCR software[3] The final recognition result is selected from the recognition results of all layers, using the following heuristic. N_v^i denote the number of valid characters (such as alpha characters and digital characters) and N_{inv}^i denote the number of invalid characters in the result of the *ith* hypothesis. We define the confidence C_i of hypothesis i as:

$$C_i = 2N_v^i - N_{inv}^i.$$

2. Remind that text may be black or white or gray.
3. we use an open OCR toolkit (OpenRTK) from Expervision

FIG. 2 – *Examples of located text in video*

3 Experiments

The whole scheme was tested on text regions located and extracted from one hour of sports video provided by the BBC, using the algorithm presented in [3]. It correctly located 98.7% text regions while providing 0.38% false alarms. We randomly selected 1208 images from the extracted text regions, mainly by time sub-sampling, providing a database of 9562 characters or 867 words. Figure 2 shows some examples. Text characters are embedded in complex background with JPEG compression noise, and the grayscale value of characters is not always the highest.

We first report results where K, the number of mixture is kept as a constant, and only the segmentation method vary (for instance, EM.2 means EM algorithm with 2 mixture or hypothesis). The character recognition rates (resp. precision rates) are computed on a ground truth basis, as the number of correctly recognized characters divided by the true total number of characters (resp. total number of extracted characters). These results are listed in table 1. It can be seen that, whatever the value of K, the GEM algorithm provides the best recognition and precision results. It is mainly due to the GEM adaptability, which, by learning the local spatial properties of the grayscale distribution, is noise adaptive and is able to better avoid over segmentation, as can be seen from the example of figure 3. Also, we can notice that the usual bimodality (K=2) hypothesis is not the best one. The explanation might be the fact that, in some instances, text images are composed of the grayscale values of characters, some contrast region around characters, and background (see figure 2).

The last rows of table 1 lists the results obtained by generating 9 hypothesis (using K=2 to 4). Even with our simple confidence criteria, the results are improved and attain a 90% character recognition rate and 86.2% word recognition rate, which constitutes a reduction of about 22% for both character and word error rate with respect to the best EM algorithm with fixed K (i.e. EM.3).

The word recognition can be improved by keeping results from two or three hypothesis with the highest confidences. We obtain a 91.2% word recognition rate using the highest two hypothesis and a 94.9% using three hypothesis in GEM, which significantly improve the result of 86.2% using only one hypothesis. This can yield better text searching results by offering more precise keywords in image and video indexing and retrieval system.

4 Conclusion

In this paper, we proposed a multi-hypothesis scheme for segmenting and recognizing embedded text of any grayscale value in image and video. Three segmentation algorithms: basic EM, MRF, and GEM are presented and compared in this scheme. The experiments show that combining the hypothesis generated by different number of Gaussians using GEM algorithm improves the segmentation and recognition performance of using basic EM and MRF with single number of Gaussians.

methods	Ext.	CRR	Prec.	WRR
EM.4	9182	83.4%	86.9%	77.9%
MRF.4	9094	84.1%	88.4%	78.7%
GEM.4	9308	87.5%	89.8%	81.8%
EM.3	9075	86.9%	91.6%	82.1%
MRF.3	9077	87.1%	91.7%	81.6%
GEM.3	9214	88.9%	92.2%	84.4%
EM.2	7535	69.1%	87.7%	59.2%
MRF.2	7803	71.8%	87.9%	61.8%
GEM.2	8788	85.4%	93.0%	82.4%
EM.4.3.2	9716	87.1%	85.7%	84.0%
MRF.4.3.2	9705	87.7%	86.4%	83.4%
GEM.4.3.2	9683	90.0%	88.9%	86.2%

TAB. 1 – *Recognition results in extracted characters (Ext.), character recognition rate (CRR), precision (Prec.) and word recognition rate (WRR).*

FIG. 3 – *Noise adaptive: Segmentation output of the present 3 algorithms and recognition results using 2 and 3 Gaussians.*

Références

[1] H. Kamada amd K. Fujimoto. High-speed, high-accuracy binrization method for recognizing text in images of low spatial resolutions. In *ICDAR*, pages 139–142, Sept. 1999.

[2] B. Chalmond. Image restoration using an estimated Markov model. *Signal Processing*, 15(2):115–129, Sept. 1988.

[3] Datong Chen, Herv Bourlard, and Jean-Philippe Thrian. Text identification in complex background using svm. In *Proceedings of the Int. Conf. on computer vision and pattern recognition*, Hawaii, USA, Dec. 12 2001. Published in Proceeding of the Int. Conf. on computer vision and pattern recognition, 2001.

[4] Datong Chen, Kim Shearer, and Herv Bourlard. Text enhancement with asymmetric filter for video ocr. In *Proc. of the 11th Int. Conf. on Image Analysis and Processing*, Sept. 2001.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum-likelihood from incomplete data via the em algorithm. *Royal Statistical Society*, B-39:1–38, 1977.

[6] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *PAMI*, Vol.6, No.6:721–741, Nov. 1984.

[7] O. Hori. A video text extraction method for character recogntion. In *ICDAR*, pages 25–28, Sept. 1999.

[8] H. Li and D. Doermann. Text enhancement in digital video using multiple frame integration. In *ACM Multimedia*, pages 385–395, 1999.

[9] R. Lienhart. Automatic text recognition in digital videos. In *Proc. SPIE, Image and Video Processing IV*, Jan. 1996.

[10] 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*, Jan. 1998. Bombay.

[11] K. Sobottka, H. Bunke, and H. Kronenberg. Identification of text on colored book and journal covers. In *ICDAR*, pages 57–63, 1999.

[12] V. Wu, R. Manmatha, and E. M. Riseman. Finding text in images. In *Proc. ACM Int. Conf. Digital Libraries*, pages 23–26, 1997.