MULTI-FRAME COMBINATION FOR ROBUST VIDEOTEXT RECOGNITION

Rohit Prasad, Shirin Saleem, Ehry MacRostie, Prem Natarajan, Michael Decerbo

BBN Technologies, 10 Moulton Street, Cambridge, MA 02138, USA

ABSTRACT

Optical Character Recognition (OCR) of *overlaid* text in video streams is a challenging problem due to various factors including the presence of dynamic backgrounds, color, and low resolution. In video feeds such as Broadcast News, a particular overlaid text region usually persists for multiple frames during which the background may or may not vary. In this paper we explore two innovative techniques that exploit such multi-frame persistence of videotext. The first technique uses multiple instances to generate a single enhanced image for recognition. The second technique uses the NIST ROVER algorithm developed for speech recognition to combine 1-best hypotheses from different frames of a text region. Significant improvement in the word error rate (WER) is obtained by using ROVER when compared to recognizing a single instance. The WER is further reduced by combining hypotheses from frame instances, which were generated using character models trained with different binarization thresholds. A 21% relative reduction in the WER was achieved for multi-frame combination over decoding a single frame instance.

Index Terms— Optical Character Recognition, Hidden Markov Models, Videotext

1. INTRODUCTION

Huge explosion of multimedia content, especially in form of archived or streaming videos has resulted in a critical need for indexing and archiving such content. While most of the information in video such as Broadcast News (BN) videos is in the visuals (faces, scenes, etc.) and the audio, often unique information is present in text form either as *overlaid* text that describes the scene or *scene* text that appears as part of the scene. A key step in indexing based on text in video is to recognize such text [1],[2]. In [2], we had presented a hidden Markov model (HMM) based recognition system configured for recognizing overlaid text in BN videos. The basic component of the system described in [2] is our script-independent Byblos Optical Character Recognition (OCR) [3] engine, which is designed for machine-printed documents and has been recently applied to handwritten documents [4].

The emphasis in [2] was on customization of the Byblos OCR system for videotext recognition. Most of the customization was focused on the pre-processing of the text regions. Since videotext typically has low resolution, the first step in pre-processing was to upsample the text images by a fixed factor. Next, given that our OCR engine expects black on white binarized images, we also developed novel techniques for binarizing the color text images.

In this paper, we postulate that the multi-frame persistence of videotext can be exploited to mitigate challenges posed by varying characteristics of videotext across frame instances. In [2], we leveraged the multi-frame persistence property of videotext to generate a single, "contrast enhanced" image for downstream processing. In this paper, we first compare the recognition performance on the contrast enhanced image to recognizing an empirically determined single best instance of a text region. Next, inspired by the system combination [5][6] approaches that have been used with significant success in speech recognition, we explore combining hypotheses from multiple instances of a particular text region. Specifically, we apply the Recognizer Output Voting Error Reduction (ROVER) algorithm [5] for combining 1-best hypotheses from multiple instances to generate a hypothesis which is significantly better than any single instance. Additional reduction in WER are achieved by: (a) incorporating the contrast enhanced image in the hypotheses set for combination, and (b) using character models trained with different binarization thresholds to decode different instances of a text region.

2. OVERVIEW OF HMM BASED VIDEOTEXT OCR

Our videotext OCR system is a customized version of the HMM based BBN Byblos OCR system developed for recognizing text in printed documents. A pictorial representation of the BBN Byblos OCR system [3] is given in Figure 1. Knowledge sources are depicted by ellipses and are dependent on the particular language or script. The OCR system components themselves are identified by rectangular boxes and are independent of the particular language or script. Thus, the same OCR system can be configured to perform recognition on any language.

The BBN Byblos OCR system can be subdivided into two basic functional components: training and recognition. Both training and recognition share a common pre-processing and feature extraction stage. The pre-processing and feature extraction stage starts off by first deskewing the scanned image and then locating the positions of the text lines on the deskewed image. Next, the feature extraction program computes a feature vector, which is a function of the horizontal position within the line. First, each line of text is

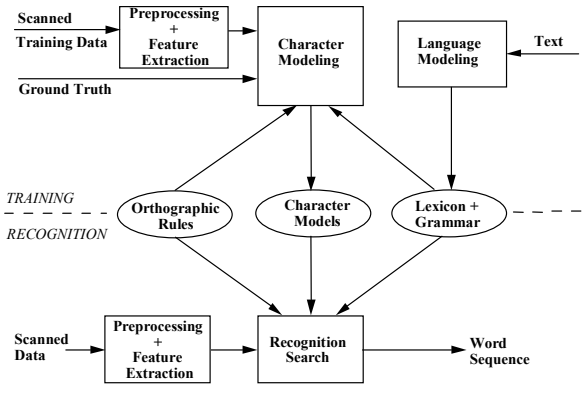

Figure 1: Hidden Markov model based Byblos OCR engine.

horizontally segmented into a sequence of thin, overlapping steps. For each frame we then compute a script-independent, feature vector that is a numerical representation of the frame.

The character models comprise of multi-state, left-to-right HMMs. Each state has an associated output probability distribution over the features. The character HMMs are trained on transcribed text lines using the Expectation Maximization (EM) algorithm. Note that the model topology including number of states, and allowable transitions is typically optimized for each script.

The language model (LM) used in the BBN Byblos OCR engine is a character or word n-gram LM estimated from the character HMM training data and other available sources of text. The recognition engine performs a two-pass search. The first pass uses a bigram LM to generate a lattice of characters or words. The second pass uses a trigram LM and optionally more detailed character HMMs to generate a 1-best hypothesis, N-best hypotheses, or a lattice.

The videotext OCR problem can be broken down into these three broad steps [2]:

1. *Text detection*: Detecting the existence and location of text within each frame in the video stream.
2. *Pre-processing*: Enhancing (removing background, upsampling, etc.) and binarizing the text image.
3. *Recognition*: Recognizing detected and pre-processed videotext.

In this paper all experiments are performed on manually detected text regions. Therefore, we focus on steps 2 and 3.

Pre-processing of videotext involves two key steps which are different from the processing of document images. The first step is to upsample the videotext region by a fixed factor (typically a factor of 4). The upsampling is performed to mitigate the effect of low resolution of videotext. The second step is to binarize the color text images into black text on white background or vice-versa, depending on the text and background characteristics. For binarization, we use a simple intensity-based procedure. In this procedure, all pixels with intensity greater than some threshold are set to black in the output image. For images with low-intensity text, pixels with intensity less than the threshold are set to black. The thresholds on the intensity are determined empirically on a development set as a percentile of intensity for each frame.

Following binarization, we extract the same set of features from videotext as for machine-printed OCR [3]. For recognition, we estimate character HMMs from the available training data with different parameter tying configuration depending on the amount of available training data. Next, the two-pass recognition strategy described earlier is used to recognize *all* I-frames for text regions in a development set. Then, we empirically determine the I-frame that results in lowest character error rate (CER) or word error rate (WER). On a validation set as well as for the runtime system, the recognition result from the empirically determined lowest CER/WER I-frame is used for evaluating performance.

3. VIDEOTEXT OCR CORPUS

The results reported in this paper are performed on overlaid videotext data collected from English Broadcast News videos. For our experiments we used the TDT-2 corpus [7] of CNN and ABC news broadcasts recorded in 1998. We annotated text region boundaries and frame spans manually. Each text region consisted of a single line of text with possibly multiple words. A single transcription ground truth value was assigned to each text region. Approximately 7 hours of video each from CNN and ABC was manually annotated. All text was annotated except for the moving text crawler in the CNN videos.

The text density in CNN was significantly higher for CNN than for ABC: 6.6 text regions per frame versus 2.1 text regions per frame. The corpus therefore contained significantly more CNN text data. Specifically, for CNN we annotated 16,719 text regions and for ABC 5,567 text regions were annotated. We held out a fair development set of 871 regions for CNN and 475 regions for ABC – none of the regions in the development set were included in the training set.

In addition to reporting results on the fair development, we present results on data provided by NIST as dry-run data for 2005 videotext OCR evaluation. The NIST dry run test set is from the same source channels as our internal set, but from a different time period in 1998. In total, there are 537 text regions in this test set.

4. MULTI-FRAME CONTRAST ENHANCEMENT

A convenient property of overlaid videotext is that the text remains relatively constant in appearance over a few frames, while the background varies. In [2], we leveraged this characteristic for improving the quality of the binarized images. For images that contain light/dark text on a dark/light background, we compute an enhanced image by taking the minimum/maximum intensity value across a number of instances of the text line, after they are aligned. This contrast enhanced image, referred to as "min-image" [2] is then binarized using the procedure described in Section 2 before training and/or recognition.

In addition to providing an enhanced image for recognition, the above procedure provides an alternative approach to the empirical selection described in Section 2 for generating the 1-best character sequence for a text region 2. That is, instead of empirically determining the best I-frame to decode on a development set, one only needs to recognize the contrast enhanced image for each text region. Therefore, we performed experiments to compare the contrast enhanced image to the empirically determined best I-frame for recognition.

First, we trained character HMMs on the entire training data consisting of 22K text regions. For each text region in the training corpus, we included 5 uniformly selected instances for character HMM estimation. This was done to increase the coverage of different types of distortions that manifest themselves over the lifetime of a text region. All training images were binarized using a threshold on pixel intensity. This threshold was chosen to be 80^{th} percentile for high intensity text and 20^{th} percentile for low intensity text. A trigram character LM was estimated from the same training data. Including the punctuations and numerals, the recognition lexicon consisted of 86 characters. Each character HMM had an associated 512 Gaussian mixtures for modeling the output feature distribution at each state.

The first row of Table 1 shows that the WER on the 5^{th} I-frame decoded with the models described above is 32.7%. Next, we generated a min-image using 15 uniformly selected I-frames for each text region in the test data. Decoding with the same models as used for recognizing the 5^{th} I-frame resulted in a WER of 32.2%, which is 0.5% absolute better than the WER on the 5^{th} I-frame. Given that the optimal model to decode the min-image is likely to be the one that is trained on min-images, we trained a new set of character HMMs on the min-images in the training corpus. Next, we decoded the min-image in the test data with these models. The WER reduced to 32.0%, a 0.7% absolute reduction in WER over the 5^{th} I-frame.

Frame(s) for Training	Frame for Reco.	%WER
5 I-frames per region	5th I-frame	32.7
Same as above	Min-image	32.2
Min-image	Min-image	32.0

Table 1: Improvements in WER on development set for using contrast enhanced text image.

5. MULTI-FRAME HYPOTHESES COMBINATION

5.1. Motivation for multi-frame hypotheses combination

The characteristics of the text (e.g., contrast) that impact recognition accuracy can change significantly from one frame instance to another. As a result, the error rate and the type of errors vary significantly across different instances of a text region. Therefore, we performed experiments to characterize the change in error rate from one instance to another by measuring the error rate associated with a single selected instance and comparing it with the error rate of the best recognition result picked (using the reference transcription) from uniformly sampled 5, 15, and 25 I-frames, respectively. For the single instance, as described in section 3, we selected the 5th (or last) I-frame for each text region.

In Table 2, we summarize the *oracle* WER for selecting the best hypothesis from different I-frames for a particular text region. The results indicate that significant improvements in WER can be achieved by *selecting* the best hypothesis from even 5 instances of a text region. Including more number of instances of a text region into the oracle selection gives further reduction in the WER, however the improvements seem to saturate after 15 instances.

Condition	%WER
Recognition on 5th I-frame (Baseline)	32.7
Oracle for 1-best across 5 I-frames	23.6
Oracle for 1-best across 15 I-frames	22.7
Oracle WER for 1-best across 25 I-frames	22.5

Table 2: Oracle WER on development set for selecting best hypothesis across multiple instances of a text region.

5.2. Multi-frame hypotheses combination using ROVER

In automatic speech recognition (ASR) it has been shown [5],[6] that a single input waveform can be processed by multiple systems and then the different system outputs can be combined to produce a 1-best answer that is significantly better than the output of any single system. Most of these combination approaches are based on the NIST ROVER [5] principle. In the case of videotext OCR we have multiple instances of the input image that are processed by the same recognition engine – *but the key hypothesis combination principle is still the same.* As shown in Table 2, the lower bound in error rates across 15 I-frames is less than two-thirds the error rate of a single-frame answer. Therefore, in this section we explore applying the NIST ROVER algorithm for combining hypotheses from multiple frame instances of a text region.

The NIST ROVER [5] algorithm has two steps. First, the multiple word hypotheses are aligned using dynamic programming (DP) into a single word transition network. Next, each branching point in the word transition network (WTN) is evaluated using a voting mechanism, where the link with highest number of votes is selected as the best scoring word. The sequence of best scoring word selected from each branching point constitutes the 1-best word sequence. The use of confidence scores for each word is optional for the voting. However, confidence scores have shown to improve robustness, especially when there are fewer system outputs to be combined.

Application of the ROVER algorithm to videotext recognition consists of two steps, which are similar to the steps in ASR system combination. First, using DP the 1-best character hypotheses from different frame instances of a text region are aligned to create a WTN for each text region. Next, the ROVER voting algorithm is applied to the WTN to generate the best character sequence.

In the first set of experiments using NIST ROVER, we compared combining hypotheses from 5 I-frames and 15 I-frames respectively. The character HMM and character LM described in the first row of Table 1 was used to generate 1-best character sequences for the I-frames of a particular text region. Confidence scores for 1-best hypothesis for each frame instance was generated using consensus network [8] transformation of character lattices produced by our two-pass OCR decoder.

As shown in Table 3, applying ROVER on hypotheses from 5 I-frames results in a 6% relative improvement over the baseline result of using the output from a single frame instance. Combining 15 I-frames resulted in further improvement and the overall reduction in WER is 8% relative to the baseline.

In Figure 2, we illustrate the effectiveness of ROVER based hypotheses combination. As shown in the figure, the error characteristics vary significantly across frame instances, primarily due to the different types of distortions in each instance. Although several instances, including the 5th I-frame contain errors, the ROVER still results in the correct answer.

Text Region Image	Recognition Result
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOOW INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOOW INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOOW INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOGY INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOOR INC
(MU) Micron Technology Inc	ASTRICASE REASEARIC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOGY INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOGY INC
(MU) Micron Technology Inc	(MU) MONON TECHNOLOOW INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOGY INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOGY INC
(MU) Micron Technology Inc	(MU) MICRON TECHNOLOGY INC
ROVER Result	(MU) MICRON TECHNOLOGY INC

Figure 2: Example of ROVER hypotheses combination.

5.3 ROVER with different binarization thresholds

In the results reported thus far, we have trained our character HMMs with 5 I-frame instances of a text region. All 5 instances are binarized at the same threshold (80th percentile of pixel intensity for high-intensity text images). Different instances of a text region in the test set are also binarized at the same thresholds. Since overlaid videotext oftentimes consists of non-uniform, noisy background, it is unlikely that a single threshold for binarization will perform the best across all instances of a text region. Therefore, we trained three sets of models on the same 5 I-frame instances of a text region in the training data using a binarization threshold of 75th, 80th, and 85th percentile, respectively.

On the test data, instead of binarizing all 15 I-frames with a single threshold, we now use the following interleaved ordering for binarizing the same set of 15 I-frames. The first I-frame is binarized with a threshold of 75h percentile, the second instance at

80th percentile, the third at 85th percentile, the fourth at 75th percentile, and so on. Each binarized frame is decoded with character HMMs trained using text images binarized with the matching threshold, that is, a frame binarized at 75th percentile is decoded with models trained with 75th percentile binarization threshold applied to the training images. Thus, the set of images we are decoding in the test set is the same as the 15-frame combination described in Section 5.2, but 10 of the 15 frames are binarized at a threshold different from 80th percentile and decoded with models trained with the matched binarization threshold (75th or 85th percentile).

As shown in Table 3, combining 15 I-frame hypotheses generated with different binarization strategy described above lowers the WER to 29.0%, a 1.2% absolute reduction in WER over combining hypotheses from the same 15 frames binarized at a single threshold and processed using a single set of character HMMs trained with the same threshold. Therefore, the multiple binarization strategy is effective in capturing the variation in the characteristics of videotext across different frame instances.

Condition	%WER
Recognition on 5th I-frame (Baseline)	32.7
Recognition on Min-image	32.0
ROVER on 5 I-frames	30.8
ROVER on 15 I-frames (same binarization)	30.2
ROVER on 15 I-frames (3 sets of 5 I-frames binarized at 3 different thresholds)	29.0
ROVER on 18 frames (Above 15 I-frames + Min-image binarized at 3 different thresholds)	27.2

Table 3: Improvements obtained using multi-frame hypotheses combination on development set.

In Section 4, we showed that the contrast enhanced image (min-image) results in a lower WER than the WER on the 5th I-frame. Although the WER reduction is modest, the results on the min-image have different error characteristics than the regular I-frames. Therefore, we decided to incorporate the min-image into our hypotheses combination framework. First, we binarized all min-images in the training corpus at three different binarized thresholds (75th, 80th, and 85th percentile). Next, we estimated separate character HMMs from each of these different sets of binarized min-images. On the test data, we binarize the min-images at the same three binarization thresholds as in training. Then, we decode these binarized images with character HMMs trained with matched binarization threshold. Finally, the 3 min-image hypotheses are added to the set of 15 I-frame hypotheses generated using different binarization thresholds and character HMMs. ROVER on this set of 18 hypotheses reduces the WER on the development set to 27.2% – a 17% relative improvement over the baseline WER of 32.7% obtained on the 5th I-frame for each text region. This reduction in error rate is about 55% of the maximum reduction possible based on the oracle WER analysis in Table 2.

Given the parameters of the ROVER algorithm, which include parameter for trading-off confidence scores and number of occurrence of a character (α), confidence score for null arcs (conf@), and the voting strategy (maximum confidence and average confidence) were optimized on the development set, we decided to compare performance on the validation set (NIST dry-run data). As shown in Table 4 below, the best multi-frame combination strategy results in a 20% relative improvement over the baseline configuration of decoding the 5th I-frame.

Condition	%WER
Recognition on 5th I-frame (Baseline)	31.2
ROVER with multiple binarization of I-frames and min-image	24.9

Table 4: Summary of improvements with multi-frame combination on NIST validation set.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we successfully applied the NIST ROVER algorithm for combining hypotheses from multiple instances of overlaid text in Broadcast News videos. We achieved large reduction in the word error rate with multi-frame combination, especially when instances of a text region are processed with different binarization thresholds. The multi-frame contrast enhancement technique showed only a modest gain over the baseline result of using the 5th I-frame. Still including the hypotheses from the contrast-enhanced image in the set of hypotheses to combine resulted in additional gain for ROVER based multi-frame hypotheses combination.

Our analysis of the results indicates that further reduction in WER can be obtained by including N-best hypotheses instead of 1-best hypothesis into the multi-frame combination. Therefore, future work will focus on confusion network combination (CNC) [9] for combining confusion networks instead of 1-best from different instances of a text region.

7. REFERENCES

[1] T. Sato et al., "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption", *ACM Multimedia Systems Special Issue on Video Libraries*, 7(5): 385-395, 1999.

[2] P. Natarajan, B. Elmieh, R. Schwartz, and J. Makhoul, "Videotext OCR using Hidden Markov Models," *Proceedings Sixth International Conference on Document Analysis and Recognition*, pp. 947 – 951, Seattle, WA, 2001.

[3] P. Natarajan, Z. Lu, I. Bazzi, R. Schwartz, and J. Makhoul, "Multilingual Machine Printed OCR," *International Journal Pattern Recognition and Artificial Intelligence, Special Issue on Hidden Markov Models in Vision*, pp. 43 – 63, 2001.

[4] P. Natarajan, S. Saleem, R. Prasad, E. MacRostie, K. Subramanian., "Multilingual Offline Handwriting Recognition," *Proceedings Summit on Arabic and Chinese Handwriting*, College Park, MD, 2006.

[5] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, CA, 1997.

[6] R. Schwartz et al., "Speech Recognition in Multiple Languages and Domains: The 2003 BBN/LIMSI EARS System," *Proceedings IEEE ICASSP*, Montreal, Canada, pp. 753-756, 2004.

[7] C. Cieri D. Graff, M. Liberman, N. Martey, S. Strassel, "The TDT-2 Text and Speech Corpus," *Proceedings DARPA Broadcast News Workshop*, 1999.

[8] L. Mangu, E. Brill, and A. Stolcke, "Finding Consensus among Words: Lattice-Based Word Error Minimization," *Proceedings EUROSPEECH*, pp. 495-498, Budapest, 1999.

[9] G. Evermann and P. Woodland, "Posterior Probability Decoding, Confidence Estimation and System combination," *Proceedings NIST Speech Transcription Workshop*, College Park, MD, 2000.