RC 22116 (99099) 17 July 2001 **Computer Science**

IBM Research Report

AUGMENTED EDIT DISTANCE BASED TEMPORAL CONTIGUITY ANALYSIS FOR IMPROVED VIDEOTEXT RECOGNITION

Hrishikesh Aradhye
The Ohio State University
140 W 19th Ave
Columbus, Ohio 43210

Chitra Dorai
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

AUGMENTED EDIT DISTANCE BASED TEMPORAL CONTIGUITY ANALYSIS FOR IMPROVED VIDEOTEXT RECOGNITION

Hrishikesh Aradhye
The Ohio State University
140 W 19th Ave
Columbus; Ohio 43210
aradhye.1@osu.edu

Chitra Dorai
IBM T.J. Watson Research Center
P.O. Box 704, Yorktown Heights
New York 10598, USA
dorai@watson.ibm.com

Abstract -

Videotext refers to text superimposed on video frames and it enables automatic content annotation and indexing of large video and image collections. Its importance is underscored by the fact that a videotext based Multimedia Description Scheme has recently been adopted into the MPEG-7 standard. A study of published work in the area of automatic videotext extraction and recognition reveals that, despite recent interest, a reliable general purpose video character recognition (VCR) system is yet to be developed. In our development of a VCR system designed specifically to handle the low resolution output from videotext extractors, we observed that raw VCR accuracies obtained using various classifiers including kernel space methods such as SVMs, are inadequate for accurate video annotation. In this paper, we propose an intelligent postprocessing mechanism that is supported by general data characteristics of this domain far VCR performance improvement. We describe Temporal Contiguity Analysis, which works independently of the raw character recognition technique and works well even for moving videotext. This novel mechanism can be easily implemented in conjunction with VCR algorithms being developed elsewhere to offer the same performance gains. Experimental results on various video streams show notable improvements in recognition rates with our system incorporating a SVM-based recognition engine and temporal contiguity analysis.

INTRODUCTION

Manual indexing of multimedia archives is an expensive and time-consuming task given the volume and richness of their content. Algorithms for automated content analysis of digital video collections are, thus, highly desirable. Text superimposed on video provides vital information about its content such as the identity of the speaker, his/her location, the topic under discussion, product captions, sport statistics; etc. Videotext extraction and recognition therefore, forms an important task in an automated video content annotation system. A study of published research in the area of videotext extraction and recognition reveals that, in spite of increased interest [1, 2, 3], a reliable general purpose

VCR system is yet to be developed. Most of the existing methods focus on a specific genre of digital video. The best reported results are those of Sato et. al. [1] who report a recognition accuracy of 73% to 89% with CNN news videos, and make program-specific assumptions such as stationary text, brightly illuminated text on a dark background, and so on. We have developed an end-to-end video character extraction and recognition system to achieve similar or better results, without making these restrictive domain-specific assumptions.

Our previous work [4] presented a scheme for automatic videotext extraction. We have recently developed the first video character recognition system [5] that employs Support Vector Machines (SVMs) as basic character classifiers on a set of 172 features specifically designed for videotext. The raw results from our SVM-based VCR system presented in Table 1 show that the range of character recognition accuracies is from 65% to 78%. The errors are due to SVM misclassifications as well as missed, merged, or broken characters during the text extraction phase. Although these raw results are unacceptable for the purpose of accurate video content annotation, they indicate a reliable baseline performance which can be improved further by postprocessing.

In this paper, we present novel postprocessing steps of videotext object synthesis, font context analysis, and temporal contiguity analysis for accurate video character recognition. The recognition accuracy is improved significantly after incorporating these steps. These individual steps can be potentially implemented in conjunction with other VCR algorithms being developed elsewhere and, in turn, lead to improved recognition performance.

SMART POSTPROCESSING FOR IMPROVED VCR

Text on video frames is intended to convey a meaningful message to the viewer. As a result, it seldom occurs as isolated characters. Additionally, temporal persistence is often employed for enhanced readability with videotext. For example, three consecutive frames with videotext are shown in Figure 1. We found that a typical text segment persists on video for at least about 20 frames. This redundancy can be exploited for higher recognition accuracy. These general videotext characteristics form the foundation of our proposed postprocessing steps.

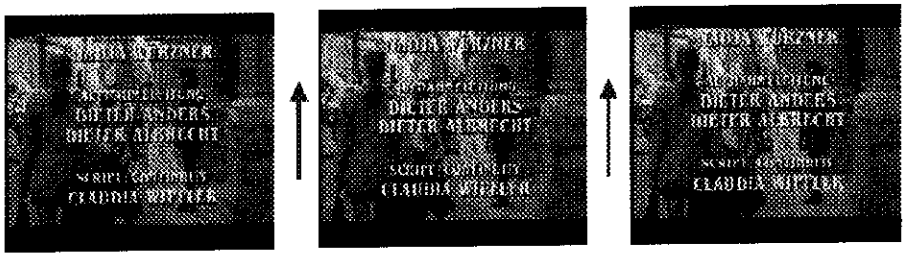

Figure 1: Temporal contiguity of videotext in digital video.

Video Character Recognition

We have developed a rich feature set for video character recognition. Our features are designed to characterize coarse shapes of the characters to handle their low resolution on video frames. We use Support Vector Machines (SVMs) as our basic classifier for character recognition. The SVM-based VCR system is trained using a One-vs-One architecture to learn the mapping from a 172-dimensional input feature space to an output recognition space consisting of 62 classes, corresponding to 26 lower case and 26 upper case letters, and 10 digits.

Object Synthesis and Font Context Analysis

Semantically and syntactically related characters are grouped together, based on the fact that videotext is rarely present in isolation. We refer to such groups as *videotext objects*. Font context analysis is performed on each text object as a first step during postprocessing. This analysis is motivated by the fact that certain pairs of characters, (e.g., 9 and g) appear very similar in images albeit very easily distinguishable by their placement above or below the baseline. We have developed a technique to determine the baseline, ascent and descent portion of a text object from the data. Let such information be referred to as the *font context*. Any recognized characters that are inconsistent with respect to the estimated distribution of the vertical space of their regions are potential errors. For each incorrect character identified in the frame, we revisit the 62-element output array of One-vs-One SVM votes. Of these 62, all class labels that are inconsistent with the estimated font context are eliminated. Among the remaining classes, the one with the maximum votes is the revised label for the video character. Table 1 shows the improved results due to font context analysis of video characters.

Temporal Contiguity Analysis

For better readability, text on video persists over multiple frames. Due to motion of the background and/or the videotext, coupled with the compression noise, the same character can be extracted/classified correctly for some of the frames and incorrectly for the others. Analysis of temporal redundancy of videotext can remove some of these errors and this is undertaken as our second postprocessing step. Temporal contiguity is typically observed in terms of stationary videotext (e.g., captions), uniform vertical motion of videotext (e.g., scrolling credits), or uniform horizontal motion of videotext (e.g., stock ticker). Other types of motion are unlikely for superimposed videotext and are not considered in this work.

Previously published literature has primarily followed the "image fusion" approach for temporal contiguity analysis [6, 7]. This approach relies on the assumptions that (i) video text is stationary and (ii) it is superimposed on a low intensity background. By considering multiple images in time, a *super-image* is constructed with reduced noise and compression artifacts for better text extraction, and hence better classification.

In this work, we propose a more general approach for multiframe fusion that handles videotext motion. For multiframe fusion, we first need to align the frames in a way that textual content is perfectly matched. Given that (a) the text could be in motion, (b) some of the characters could be missed or misclassified, and (c) the same letter could be present multiple times on the same frame and potentially in the same text object, the task of simultaneous alignment of videotext across multiple frames is non-trivial. We propose a new method for *dynamic alignment* of multiple video text objects from successive frames. These aligned video text strings could be then fused using the conventional "image fusion" or "feature fusion" approaches. In this work, we propose a complimentary approach that consists of class label fusion instead.

Multiframe Videotext Alignment Using Enhanced Edit Distance

The *Edit Distance* metric [8] provides an efficient method to simultaneously align multiple strings by penalizing three unit operations of *change*, *delete*, and *insert*. The resultant optimization problem can be solved using Dynamic Programming.

The first step of our alignment algorithm is to determine the best matches to any given videotext object across frames in either direction in time. We use an enhanced 3-D Edit Distance based Dynamic Programming approach to align these three strings. Conventional Edit Distance metric assigns equal costs to all three unit operations. The *change* operation has a cost of unity in case of mismatch of characters, and zero cost for perfectly matching characters. For our problem, we observed that such a *crisp* cost function may not be the best. Instead, we propose a classifier specific *fuzzy* cost for the *change* operation, such that the cost of mismatch is proportional to the probability of the corresponding misclassification by the video character recognizer. The misclassification probabilities are, obviously, classifier-dependent. They can be estimated from the confusion matrix generated by the particular classifier in use for video character recognition.

We also introduce a new unit operation for the augmented Edit Distance metric that is unique to videotext, namely *move*. Three constraints about the motion of videotext can be applied to these characters: (i) The videotext motion is along either the horizontal direction or the vertical direction. (ii) All characters belonging to the same text object undergo the same motion from one frame to the other (e.g., the various characters from the words "CLAUDIA WIPFLER" from Figure 1). In other words, their motion vectors are parallel and have the same magnitude. (iii) Videotext does not accelerate or decelerate, i.e., the motion vectors of the same character between two consecutive frames are parallel and have the same magnitude. For instance, the videotext from Figure 1 is seen to move with a uniform vertical velocity from frame to frame.

If these constraints are not obeyed, our augmented Edit Distance assigns a very large cost due to the unit operation *move*. If all three constraints are satisfied, the cost is 0. We thus force the dynamic programming optimization that used this enhanced Edit Distance to obey the motion constraints. At the

end of the optimization, we have an optimal alignment of the 3 strings under consideration and a measure of the corresponding alignment cost, namely the augmented Edit Distance. This measure is used to prune falsely extracted, non-text characters.

Multiframe Fusion Using Bayesian Inference

We are now in a position to fuse the aligned text objects from three temporally contiguous frames. For matches that involve characters from two or all three strings, Bayesian inference is used to conclude about the true class label by treating the VCR results as instances of a random variable and maximizing its posterior probability. The information from the confusion matrix is used for calculating the conditional probabilities. A priori probabilities for all classes are assumed to be equal. They also can be determined by modeling the class probabilities based on a dictionary or word usage in a video domain.

In general, the temporally fused string at any time instant will include less errors than the corresponding recognition results after font context analysis. While calculating the temporal fusion at time t, we can substitute the videotext from frame at $t-1$ by the temporally aggregated string at time $t-1$. For videotext that persists over a large number frames such a resubstitution approach improves the recognition accuracy by about an additional 1%.

EXPERIMENTAL RESULTS

Temporal contiguity analysis is thus seen to have a significant impact on the overall accuracy. Table 1 presents the results of our end-to-end VCR system with postprocessing on several MPEG-1 video streams. With this contiguity analysis, the false text blob extraction rate was reduced from 15% to nearly zero.

Table 1: Performance of SVM-based VCR system with smart postprocessing.

Test Video	CNN-1	T5	Bangkok
Type	News Captions	Scrolling Credits	Subtitles
Raw SVM Accuracy	(67%)	(65%)	(78%)
+ Font Context Analysis	72%	68%	87%
+ Temporal Analysis	80%	76%	89%

The processing time per frame with our algorithm varies with the number of characters on each frame. In our experiments, the average processing time was of the order of 5 fps on a Pentium III 550 MHz machine running Microsoft Windows NT. The SVM recognition required a bulk of the processing time, with the extraction and postprocessing steps taking only about 5% of the total computation.

CONCLUSION

This work presents a novel end-to-end system for the task of videotext recognition. The core of this system is a Support Vector Machine classifier. SVMs outperformed other classification algorithms, such as a Naive Bayes classifier, when tested on generated characters. We also propose a set of novel post-processing steps that significantly improve the recognition performance. Our method for font context analysis exploits the syntactic information, such as the placement of characters relative to the baseline, to report performance gains. In addition, this work also provides a new method for alignment of videotext, possibly in motion, from temporally contiguous video frames using an enhanced Edit Distance metric. Our classifier-specific label fusion scheme combines the information from the aligned text objects and Bayesian inference based on the confusion matrix for the VCR classifier. Our system demonstrates a comparative performance on several video streams to the methods presented in the literature, without making restrictive assumptions such as strong contrast and lack of videotext motion.

References

[1] T. Sato, T. Kanade, E. Hughes, M. Smith, and S. Satoh, "Video OCR: Indexing news libraries by recognition of superimposed caption," *ACM Multimedia Systems*, vol. 7, no. 5, pp. 385–395, 1999.

[2] R. Lienhart and W. Effelsberg, "Automatic text segmentation and text recognition for video indexing," *ACM/Springer Multimedia Systems*, vol. 8, pp. 69–81, January 2000.

[3] H. Kuwano, Y.Taniguchi, H. Arai, M. Mori, S. Kurakake, and H. Kojima, "Telop-on-demand: Video structuring and retrieval based on text reocgnition," in *International Conference on Multimedia and Expo*, (New York City, NY), pp. 759–762, July 2000.

[4] J.-C. Shim, C. Dorai, and R. Bolle, "Automatic text extraction from video for content-based annotation and retrieval," in *Proc. International Conference on Pattern Recognition*, vol. 1, (Brisbane, Australia), pp. 618–620, August 1998.

[5] C. Dorai, H. Aradhye, and J.-C. Shim, "End-to-end videotext recognition for multimedia content analysis," in *Proc. International Conference on Multimedia and Expo, To Appear*, 2001.

[6] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," in *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[7] T. Sato, T. Kanade, E. Hughes, and M. Smith, "Video OCR for digital news archives," in *IEEE International Workshop on Content-Based Access of Image and Video Databases*, January 1998.

[8] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms*, ch. 16, Dynamic Programming. The MIT Press, 1990.