CONFUSION NETWORK BASED VIDEO OCR POST-PROCESSING APPROACH

Anan Liu[1,2,3,4], *Jinghao Fei*[2,3,4], *Sheng Tang*[4], *Jianping Fan*[3,4], *Yongdong Zhang*[4], *Jintao Li*[4], *Zhaoxuan Yang*[1]

1. School of Electronic Engineering, Tianjin University, Tianjin, 300072, China
2. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
3. Shenzhen Institute of Advanced Technology, Shenzhen, 518054, China
4. Institute of Computing Technology, CAS, Beijing, 100080, China

liuanan@cs.cmu.edu

ABSTRACT

The paper originally presents a confusion network based framework for Video OCR post-processing. The framework consists of four parts: selection of reference and hypotheses, construction of confusion network, decoding for final output, and a novel metric of quantitatively evaluating Video OCR post-processing approaches. By integrating both visual and textual information, we construct the character transition network to reduce the error rate for OCR outputs. The large-scale experimental results demonstrate that this approach can significantly improve the accuracy of Video OCR results with only little incremental time. Moreover, with comparison and the detailed analysis, we conclude that "Voting+2-gram" is the most applicable method for real application.

Index Terms— Confusion Network, Video OCR, Post-processing

1. INTRODUCTION

Because video text is directly related to semantic information, Video OCR attracts much attention from researchers for multiple applications, such as video indexing, retrieval and content filter. However, the real application shows the dilemma of the state-of-the-art techniques between accuracy and speed. On one hand, although many heuristic methods presented in [1] show acceptable speeds, the improvement of accuracy is insignificant, especially for the complicated background and low resolution. On the other hand, even if statistic-based methods, e.g., text identification with MRF in [2], can improve the results to some extent, these complicated algorithms involve high computational cost. The situation indicates that we need to balance the effectiveness and complexity of the algorithm for the real application.

Motivated by the research in Automatic Speech Recognition (ASR) [4][5] and Machine Translation (MT)[6][7][8], we utilize the characteristics of both video and text for Video OCR post-processing to improve OCR results. From the viewpoint of video, temporal information can benefit OCR results. The existing method of using multi-frame information in [3] is to enhance the contrast and facilitate text segmentation by fusing the successive frames with the same text information. However, the improvement is trivial and the method ignores that the OCR output of each frame can be fused for a better result. From the viewpoint of text, OCR results can be modified by context information. Statistical language models, syntactic principles and so on in [6][7] have been widely used in post-processing for ASR and MT. Therefore, they can be used to improve OCR results.

In this paper, we originally propose a confusion network based framework for Video OCR post-processing which is seldom mentioned in the literature. Compared with the post-processing methods used in ASR and MT, the proposed approach fuses OCR outputs (hypotheses) of consecutive frames with the same text information as the reference to construct confusion network and implements statistical language models to modify OCR outputs. By integrating visual and textual processing, OCR result is significantly improved with only little incremental time. The main contribution of our work is to propose the framework by bringing in ASR and MT methods and fusing multimodal information for the real application and the innovation lies in each part of the framework: (a) The reference of confusion network is chosen by recognizing the fused image of consecutive frames in visual modality; (b) Character transition network (***CTN***) is constructed based on progressive multiple alignment algorithm; (c) Log-line model is formulated as the metric for choosing the best output path; (d) A new method is proposed to quantitatively evaluate Video OCR post-processing approach.

The remainder of the paper is organized as follows. In Section 2, we describe the proposed framework in detail. Then, experimental results are presented in Section 3. At last, the conclusions and future work are stated in Section 4.

2. FRAMEWORK DESCRIPTION

From Fig.1 (b), we can see that there are always missing or error characters in the OCR outputs due to multiple reasons, such as complicated background and low resolution.

Therefore, we propose character-level confusion network based framework for Video OCR post-processing to yield a reduced error rate for OCR results. The proposed framework contains four key ingredients: reference (***Ref***) & hypotheses (***Hyp***); construction of character transition network (***CTN***); decoding for confusion output, and the metric of quantitative evaluation. The flowchart of the method is shown in Fig.2.

2.1. Reference and Hypotheses

It is stated in [4][8] that a carefully chosen alignment reference is important for a good performance. Based on video characteristics, it is perceptive that the OCR output of fused image of consecutive frames with the same text information is usually better than those of individuals. Therefore, we choose the OCR output of the fused image as the alignment reference and the OCR outputs of individuals as hypotheses. For the detailed methods of multi-frame fusion, test segmentation and test recognition, people can refer to [3].

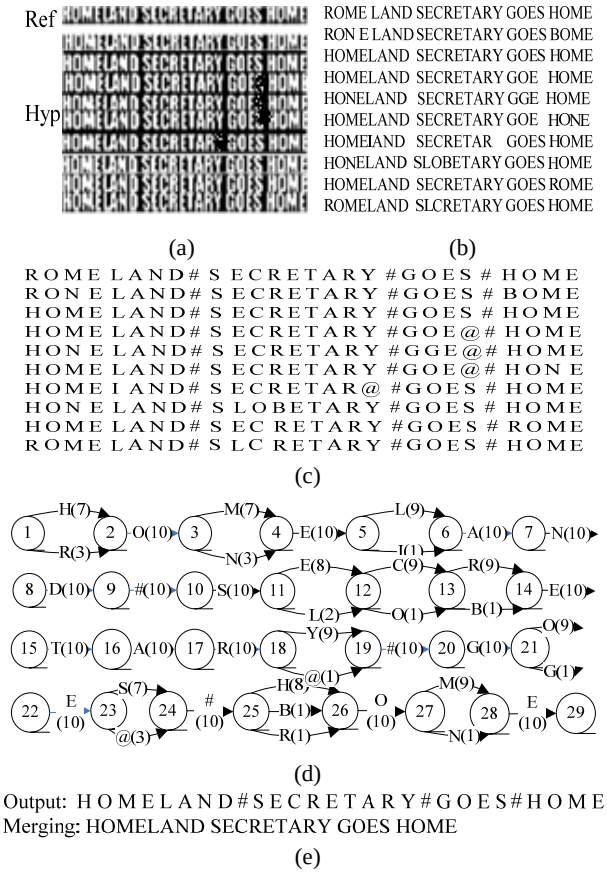


Fig.1 CTN construction. (a) OCR inputs: Text segmentation results; (b) OCR outputs for Ref and Hyp; (c) Preprocessing ("#" for blank, "@" for Null); (d) CTN construction; (e) Decoding for confusion output.

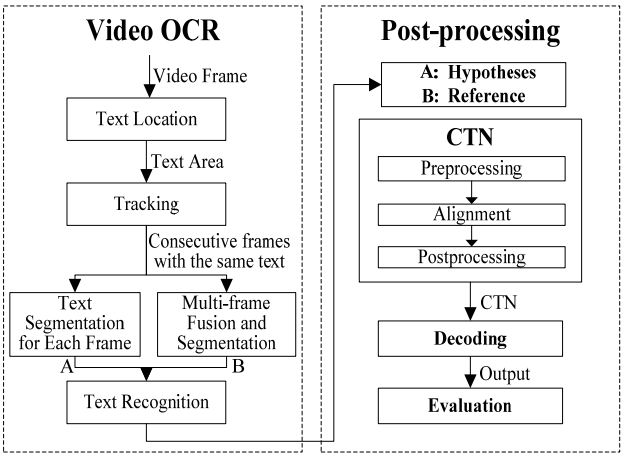


Fig.2 Flowchart of confusion network based framework for Video OCR post processing.

2.2. Construction of Character Transition Network

Character Transition Network (***CTN***) is constructed with the following three steps.

2.2.1. Preprocessing

OCR outputs usually comprise incomplete words with missing characters. For ***CTN*** construction, we must divide the words into characters with separating symbol, "#", shown in Fig.1 (c).

2.2.2. Alignment

Character alignment is the key problem for ***CTN*** construction. We implement the progressive multiple alignment method for its higher effectiveness. The algorithm is introduced in the biological sequence literature [9] as follows:

1. Compute the edit distance scores and their profiles for each of the *N(N-1)/2* pairs of strings;
2. Repeat the following until one profile remains
   (a) Select the profile for the least edit distance string-string, string-profile or profile-profile pair;
   (b) Compute the edit distance between the selected profile and the remaining strings and profiles.

The result of the algorithm is a tree structure and the similar strings appear closer at the leaf level.

2.2.3. Postprocessing

After alignment, the same characters aligned together are merged so that ***CTN*** only comprises a sequence of unique characters. Moreover, each character is assigned a score by voting scheme. A constructed ***CTN*** is shown in Fig.1 (d).

2.3. Decoding for Confusion Output

As for decoding for confusion output, we propose the log-line model with multiple effective knowledge sources for choosing the optimal path. The first one is character frequency (***P**_{CF}*) which denotes the probability of character existence. It is calculated by the negative logarithm of the probability. The second source is statistical language model which means the context information. Character-level N-gram language models (***P**_{LM-N}*) can be trained with lexicon. Then, we formulate the log-line model as follows:

$$\hat{O} = \arg\ \min(-\lambda_c * \log P_{CF} - \sum_{i=1}^{N} \lambda_i * \log P_{LM-i}) \quad (\lambda_c + \sum_{i=1}^{N} \lambda_i = 1) \quad (1)$$

The equation means that with character frequency and language models, we can decode the confusion network by choosing the shortest path. After getting the confusion output, we can delete "#" and merge the characters into words as it is shown in Fig. 1(e).

2.4. Evaluation Method

Motivated by Bilingual Evaluation Understudy in [10], we propose the evaluation method for the improvement of the proposed Video OCR post-processing approach without heavily manual effort for the computation of accuracy.

We use confusion output as the reference and OCR results of individual frames as hypotheses. First, we compute the n-gram matches between the reference and each hypothesis. Then, we compute the accuracy score, ***P**_n*,, which indicates the character choice and order, by the ratio of the number of the clipped n-gram for all hypotheses to the number of candidate n-grams in all the hypotheses.

In addition, we bring in the penalty factor, ***F***, which represents the factor of length. It is obvious that the OCR results of hypotheses longer or shorter than the reference should be penalized. We formulate the factor between ***i**th* hypothesis and reference, ***F**_i*, as follows:

$$F_i = \exp(-|1 - L_{h-i} / L_r|) \quad (2)$$

where ***L**_{h-i}* and ***L**_r* respectively denotes the length of ***i**th* hypothesis and reference.

Then, we formulate the improvement of post-processing, ***I***, as follows:

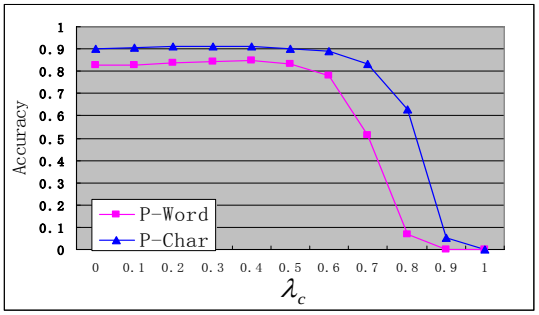
$$I = [\exp(\sum_{n=1}^{N} w_n \log p_n)] * [\sum_{i=1}^{M} w_i F_i] \quad (\sum_{n=1}^{N} w_n = 1, \sum_{i=1}^{M} w_i = 1) \quad (3)$$

where ***N*** denotes the length of n-gram from 1 to N; ***M*** denotes M hypotheses; ***w**_n* and ***w**_i* means the positive weights.

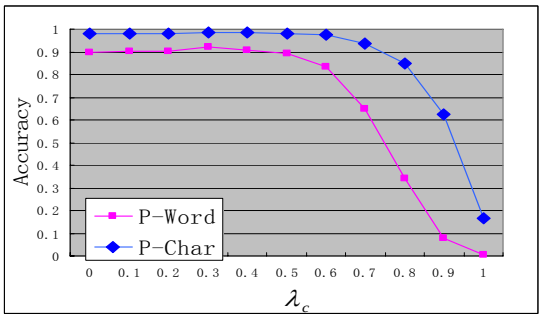
3. EXPERIMENTAL RESULTS

For evaluation, we select 50 video clips, totally 1734 frames (size: 352*240) from CNN and NBC news videos in the development data of TRECVID 2006 as the test data.

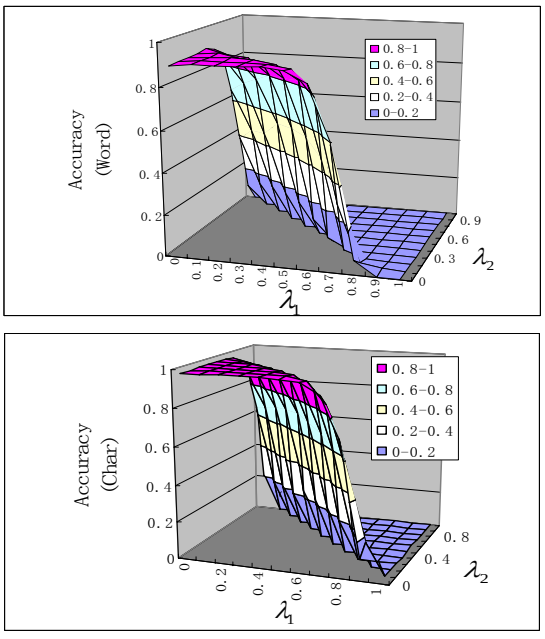
We implement the method in [3] on our test data and the result is considered as the baseline. Besides, Character-level n-gram language models are trained with the lexicon of Kingsoft Powerword, a famous dictionary software. With these language models and the variation of λ_c and λ_1, there are different rules for decoding. Considering both accuracy and speed, we respectively implement "Voting (V)" (λ_c=1, λ_1=0), "Voting+1-gram (V+G1)" (λ_c=0.4, λ_1=0.6), "Voting + 2-gram (V+G2)" (λ_c=0.3, λ_2=0.7),



(a) Relationship between accuracy and λ_c for "Voting+1-gram"



(b) Relationship between accuracy and λ_c for "Voting+2-gram"



(c) Relationship between accuracy and λ_1, λ_2 for "Voting+1-gram+2-gram"

Fig.3 Parameters decision for different decoding rules.

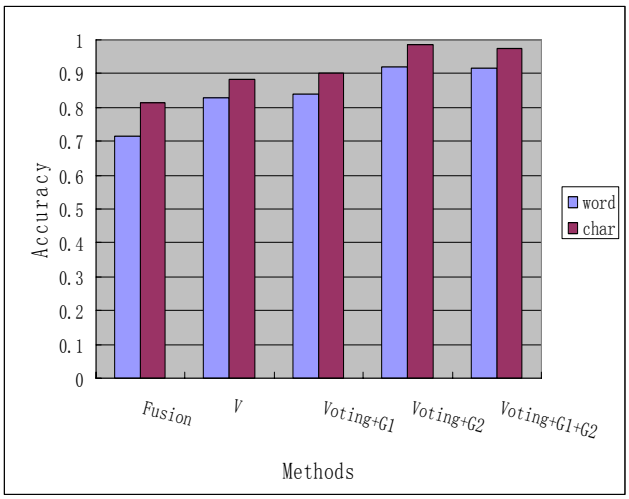


Fig.4 Comparison of experimental results.

Table 1 Comparison of consuming time for post-processing

Method	V	V+G1	V+G2	V+G1+G2
Time(ms/f)	1.07	1.36	1.77	1.92

(Note: About 20 characters per frame; Pentium® 4 at 2.8 GHz, 512M; About 250 ms/f for baseline)

"Voting+1-gram+2-gram (V+G1+G2)" (λ_c =0.6, λ_1=0.1, λ_2=0.3) for test. The parameters are chosen based upon the corresponding experimental results shown in Fig.3. From the comparison of experimental results in Fig. 4 and Table 1, the detailed analysis is presented as follows:

(a) The accuracy of multi-frame fusion method is relatively low and does not satisfy the real application..

(b) With the knowledge of character frequency, accuracy is obviously improved. This result indicates that the OCR outputs of consecutive frames with the same text information can be used for reducing character error rate. At the same time, the incremental time for this operation is only 1.07ms/f which is much less than that of baseline.

(c) Although both voting and 1-gram language modal are used, accuracy is only improved insignificantly compared to that of voting method with more consuming time. In fact, 1-gram language modal means the character frequency in lexicon. Due to the similar roles both factors play, the accuracy of this method of voting plus 1-gram modal is not obviously improved.

(d) Compared to the methods above, the method of V+G2 greatly improves the accuracy with acceptable consuming time because of two reasons. On one hand, voting is used for reducing character error rate. On the other hand, 2-gram modal provides context information for improvement. Therefore, this method effectively improves the accuracy.

(e) It is not always true that the more language modals are used, the higher accuracy will be got. The accuracy decreases slightly when both 1-gram and 2-gram modals are used on the basis of Voting. Voting reflects the accuracy with the local information, 2-gram provides the context information and 1-gram is related with both aspects. The tradeoff between them may lead to the decrease to some extent. Therefore, not only the worse result is got compared to that of "V+G2" but also more time is consumed.

4. CONCLUSION AND FUTURE WORK

In this paper, we originally propose confusion network based framework for Video OCR post-processing. With ***Ref*** and ***Hyp*** selection, ***CTN*** construction and decoding, we construct character-level ***CTN*** to improve the accuracy of Video OCR. With the proposed evaluation metric, the large-scale experimental results demonstrate that this approach can significantly improve the accuracy of Video OCR results with only little incremental time. Considering both accuracy and consuming time, "V+G2" (λ_c=0.3, λ_2=07) is the best method for confusion network based video OCR post-processing approach.

In the future, our research will focus on the novel decoding method to improve OCR results for the real application.

5. REFERENCES

[1]. Keechul Jung, Kwang In Kim, Anil K. Jain, "Text Information Extraction in Images and Video: A Survey," Pattern Recognition, Vol. 37, Issue. 5, pp. 977-997, 2004.

[2]. Yefeng Zheng, Huiping Li, David Doermann, "Text Indentification in Noisy Document Images Using Markov Random Field," In: ICDAR 2003.

[3]. Rainer Lienhart, Axel Wernicke, "Localizing and Segmenting Text in Images, Videos and Web Pages," IEEE Trans. Circuits Syst. Video Technol., Vol. 12, No. 4, pp. 256-268, 2002.

[4]. J. G. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," in Proc. IEEE ASRU Workshop, 1997.

[5]. H. Schwenk, J. L. Gauvain. Improved ROVER using language model information [A]. In: ISCA ITRW, 2000.

[6]. Antti-Veikko I. Rosti, Necip Fazil Ayan, et al, "Combing Outputs from Multiple Machine Translation Systems," In: HLT-NAACL, 2007.

[7]. Kenji Yamada, Kevin Knight. A syntax-based statistical translation model. In: ACL, 2001.

[8]. K.C. Sim, W. Byrne, et al, "Consensus Network Decoding For Statistical Machine Translation System," In: ICASSP, 2007.

[9]. R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press, 1998.

[10]. Kishore Papineni, Salim Roukos, et al, "BLEU: a Method for Automatic Evaluation of Machine Translation," In: ACL, 2002.