ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096063

WL-MSR: WATCH AND LISTEN FOR MULTIMODAL SUBTITLE RECOGNITION

Jiawei Liu[1,2], *Hao Wang*[1,2], *Weining Wang*[1], *Xingjian He*[1,2], *Jing Liu*[1,2*]

[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
{liujiawei2020, wanghao2019}@ia.ac.cn, {xingjian.he, weining.wang, jliu}@nlpr.ia.ac.cn

ABSTRACT

Video subtitles could be defined as the combination of visualized subtitles in frames and textual content recognized from speech, which play a significant role in video understanding for both humans and machines. In this paper, we propose a novel **W**atch and **L**isten for **M**ultimodal **S**ubtitle **R**ecognition (WL-MSR) framework to obtain comprehensive video subtitles, by fusing the information provided by Optical Character Recognition (OCR) and Automatic Speech Recognition (ASR) models. Specifically, we build a Transformer model with mask and crop strategies and multi-level identity embeddings to aggregate both the textual results and features of the two modalities. To pre-filter out the noise items in OCR results before fusion, we adopt an OCR filter based on ASR results and confidence scores of OCR. By combining these techniques, our solution wins the 2nd place in Multimodal Subtitle Recognition Challenge on ICPR2022.

Index Terms— Video Subtitle Recognition, Multimodal Fusion, Speech Recognition, Transformer

1. INTRODUCTION

Nowadays, video has become one of the most prevalent information carriers. Video subtitles, which contain what people are talking about, are important for video understanding, especially for human-centric videos. ASR models [1, 2] could be used to recognize what is said in the videos, but the difficulty lies in the situations involving homophones, dialects, or environmental noise. Other methods for recognizing video subtitles mostly rely on OCR models to detect [3, 4] and recognize [5, 6] texts from visual frames. Nonetheless, as in Fig. 1, there are noisy texts besides the actual subtitles or no visualized subtitles in certain videos. In this paper, we focus on the task of Multimodal Subtitles Recognition (MSR), which requires extracting video subtitles considering both visual frames and speech, and solve the above problems. MSR is indeed a challenging and practical task, as shown in Fig. 1.

There are three main issues that make MSR a challenging task: (1) Multimodal information from vision and speech must be used and fused effectively. In fact, due to the modality complementarities of vision and speech, multimodal fusion can solve the aforementioned problems of single modality methods. (2) Too many noisy textual items in the visualized subtitles make it more difficult to build cross-modal associations, so pre-filtering is very important. (3) It is very likely that the subtitles from a certain modality cannot be accurately obtained due to excessive noise or modality absence, as shown in Fig. 1. Therefore, it is necessary to prevent model from relying too much on vision or speech.

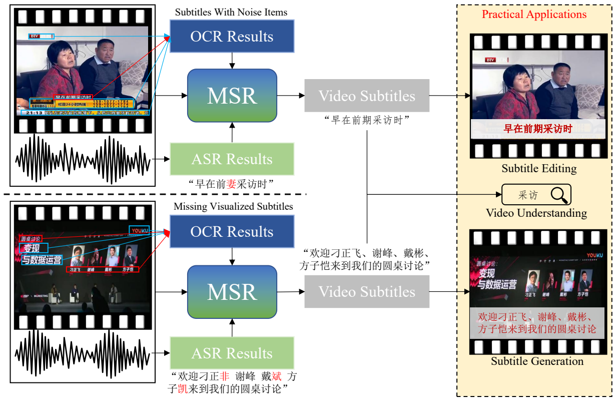

Fig. 1. An illustration of MSR and some application scenarios, i.e. subtitle editing, video understanding (such as video retrieval) and subtitle generation.

To address the above issues, we propose a novel multimodal fusion framework, named **W**atch and **L**isten for **M**ultimodal **S**ubtitle **R**ecognition (WL-MSR). An overview of the proposed framework can be found in Fig. 2. For a specific video, the Voice Activity Detection (VAD) [7] module is first used to clip it into segments. The results and features of both ASR and OCR in each segment are extracted offline using pretrained model. To pre-filter out the noisy items in visual frames, we propose an OCR filter module based on ASR results and prior knowledge. An OCR merge module is then used to remove duplicate OCR results. Results and features of ASR and OCR are further preprocessed and input to the multimodal fusion transformer model to build cross-modal associations. For the OCR results and features, multi-

[*]Corresponding author

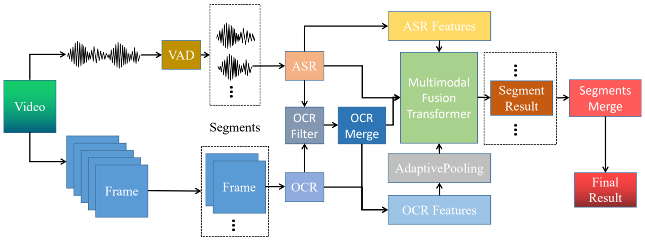

Fig. 2. Overview of our proposed WL-MSR framework for Multimodal Subtitle Recognition.

level identity embeddings, including frame id embedding, bounding box embedding, and confidence score embedding, are utilised for better location representation and credibility guidance. To reduce the dependence of the fusion model on a single modality, we introduce mask and crop strategies in the autoregressive training process, which significantly improves the generalization of the model.

Our method follows the unified multimodal direction but is designed to address several specific issues for the MSR task. We also hope that it can provide some inspiration for other applications. The main contributions are summarised as follows: (1) We propose WL-MSR, a novel multimodal fusion framework for fusing ASR and OCR results and features for video subtitle recognition. (2) Novel multi-level identity embedding for OCR results, as well as mask and crop training strategies, are introduced for the multimodal fusion model. (3) Our method has achieved excellent results in the Multimodal Subtitle Recognition Challenge, significantly outperforming the single modality methods.

2. RELATION TO PRIOR WORK

In this section, we will briefly review previous ASR-based and OCR-based methods for video subtitle recognition.

ASR-based methods are usually used for video subtitle synthesis [8], because they only use speech as input. Current end-to-end deep learning based ASR methods have shown excellent results, such as WeNet [2] and CIF [1], especially with the support of large-scale pre-training models [9]. However, the ASR methods may still fail to recognize homophones, dialects or at noisy scenes. In this work, we adopt the idea of multimodal fusion to correct those wrong characters.

OCR-based methods consist of text detector and text recognizer. Recent text detection methods focus on text of arbitrary shape in complex environments, such as segmentation-based FCENet [4]. Actually, subtitles are usually horizontal rectangles, so the horizontal text detection method [10] or simple object detector [3], is more suitable for subtitle detection. General OCR methods could be used as the text recognizer, while Connectionist Temporal Classification (CTC) [11] based models [5] are more suitable than attention-based methods [12] due to computational efficiency. Nonetheless, the OCR-only method still has the problems that removing complicated noise items only by vision is difficult and it is possible to recognize wrong words on a complex background.

3. METHOD

As shown in Fig. 1, video subtitles recognized only by OCR or ASR models have disadvantages due to noise interference. Note that visual and audio are complementary modalities, which means that the fusion of ASR and OCR could compensate for the shortcomings of single modality methods. Thus, in this paper, we cover the MSR task by fusing ASR and OCR information, and propose **W**atch and **L**isten for **M**ultimodal **S**ubtitle **R**ecognition (WL-MSR) framework, as shown in Fig. 2. Specifically, VAD [7] is used to clip the original video into segments of about 10 seconds. Then the pretrained ASR and OCR models are used to extract features and textual results, which will be further input to the proposed multimodal fusion transformer and obtain video subtitles.

3.1. OCR Filter and Merge

It should be noticed that the output of OCR is full of noise, i.e. not all detected texts are subtitles. In this case, we propose the OCR Filter, composed of a bag of matching rules based on ASR results and prior knowledge of OCR recognition, as in Fig. 3. Specifically, we calculate the **N**ormalized **E**diting **D**istance (NED) between all OCR results in a frame and the ASR results of the corresponding segment:

$$NED(t_{o[i]}, t_a) = \frac{edit\_distance(t_{o[i]}, t_a)}{max(len(t_{o[i]}), len(t_a))} \tag{1}$$

where $t_{o[i]}$ and t_a represent the i-th OCR result and the ASR result, respectively. Then the OCR item with the smallest NED is taken as the visual subtitle. Note that there may be time deviation between the subtitle display and speech. Thus, if the minimum NED is bigger than the preset threshold θ, the ASR results of the previous segment and the latter segment are used for re-matching. However, it is possible that the NED is still bigger than θ because of poor ASR results. In this case, the OCR item with the highest average confidence score is selected, as we observe that the items with high confidence are more likely to be subtitles, based on the statistics of the training data. Finally, the OCR results in a segment will be merged into a whole sentence based on their similarities.

3.2. Multimodal Fusion Transformer

In order to effectively fuse the two modalities of vision and speech, we build a **M**ultimodal **F**usion **T**ransformer (MFT) based on the vanilla transformer [13] architecture of encoder-decoder framework. In Fig. 4, we show an illustration of the training process of MFT. The text results and features obtained by ASR and OCR models are used as the inputs of the encoder, and the decoder will generate final subtitles after

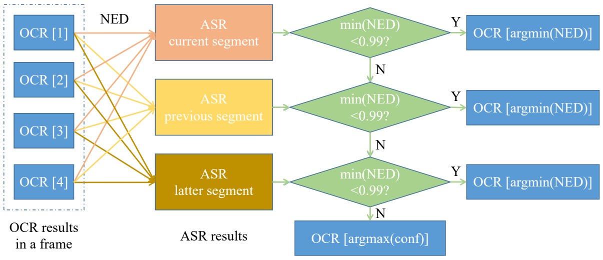

Fig. 3. An illustration of the OCR Filter, where NED is the normalized edit distance and conf is the average confidence score of an OCR result sentence.

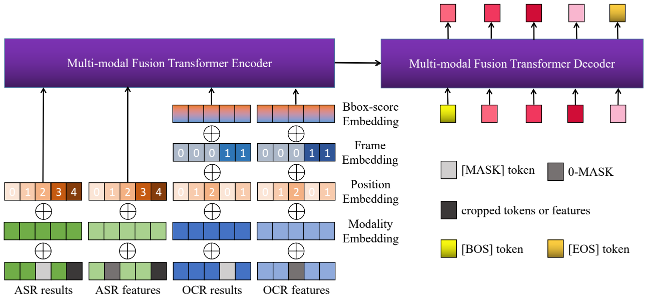

Fig. 4. Demonstration of the training process of MFT. [BOS] and [EOS] indicates begin and end of sequence.

fusion. Specifically, the sequences of different modalities are concatenated with additional respective modality embeddings and identity embeddings. The identity embeddings of ASR text results and features are learnable position embeddings of the word index. We propose multi-level identity embedding for OCR results and features as for the complex location of visual subtitles. The word index, frame index, location of bounding boxes and the confidence score of each word are all considered. The ASR annotations are adopted as the output targets, as there is less noise. We train our MFT model in an auto-regressive way, i.e., left-to-right prediction.

A key problem in training MFT is that the model will be excessively dependent on the ASR results, especially when the ASR model overfits on a small dataset. Therefore, we introduce a mask mechanism as in previous transformer-based pre-train methods [14, 15, 16]. Besides, it is observed that the voice ends before a sentence is finished in some videos. Thus, we further introduce a crop strategy for ASR results. ASR text results and ASR features at a rate of M_a are randomly masked, replaced by a specific [MASK] token and constant of 0 respectively. A sub-sequence is further taken from the end of ASR results and features with a rate of C_a, and then randomly truncated along with padding and change of attention mask. We also utilize the mask strategy on OCR results and features with a rate of $M_o = M_a \times 0.1$ for robustness.

3.3. VAD-based data augmentation

The VAD module is used in our work to clip videos into segments. We build it based on a variant of pyannote [17, 7], which contains three parts. Firstly, we use the original pyannote segmentation model to obtain segmentation scores for every time step. Then we traverse the scores and start recording a segment when the score is higher than 0.5 until the score is lower than 0.5. Finally, we clip the segments longer than 10 seconds at the time step with the smallest score.

Besides, VAD is also used on the training set to re-clip videos for data augmentation, where the annotations are obtained according to the text lengths and time steps. In this way, the augmented set not only expands the dataset but also introduces noise into the training data and narrows the distributions of the training set and validation/test set.

4. EXPERIMENTS

4.1. Dataset and Evaluation

We train and evaluate our proposed WL-MSR on ICPR 2022 Multimodal Subtitle Recognition Challenge [18] subtask3 dataset , called ICPR2022_MSR for brevity. The ICPR2022_MSR dataset contains 75 hours of video content, among which 50/5/20 hours are used for training, validation, and testing, respectively. The annotations of the training set consist of correct ASR annotations and weak OCR annotations, where all textual items in the video frames are labeled. For the evaluation, subtitles in both visual frames and speech should be recognized and merged into a whole sentence. The **C**haracter **E**rror **R**ate (CER) is used as the evaluation metric.

4.2. Implementation Details

The ASR model is implemented by a CIF-based [1] pretrained Wav2Vec2.0 (XLSR) [9], using fairseq [19]. The model is finetuned on AISHELL1 [20], ICPR2022_MSR and the augmented dataset, with a combination of cross entropy loss, quantity loss, and CTC loss as in CIF [1].

We adopt FasterRCNN [3] trained on the ICDAR [21] dataset as the text detector for OCR. The text recognizer is implemented by CRNNNet [5] trained using CTC loss. For the training data, we randomly select and combine 30k corpus from the annotation of ChineseOCR [22] and ICPR2022_MSR. After that, we paste those corpora onto the background images extracted from the videos of the ICPR2022_MSR dataset to synthesis OCR training data.

The NED threshold θ in OCR filter is set to 0.99. The MFT model is built on the seq2seq Transformer-base, with 6 encoder layers and 6 decoder layers, each with 8 heads for self-attention and cross-attention layers. The dimension of the hidden embedding is set to 768. The word embedding layer is shared between the encoder and decoder, with a vocabulary size of 3490 and an additional CELU [23] activation function and dropout probability of 0.5. The Adam

Table 1. Comparison with single modality methods on ICPR2022_MSR validation set. † indicates training on additional augmented dataset.

Method	Modality	CER
CIF [1]	ASR	0.3650
FasterRCNN[3] + CRNNNet[5]	OCR	0.2921
WL-MSR	Multimodal	0.2301
WL-MSR†	Multimodal	**0.1931**

Table 2. Final results on ICPR2022_MSR test set.

Team	CER	rank
mys, baseline_on_baseline etal.	22.20	4
flames etal.	18.26	3
alpaca, reyne, etal.	**12.96**	1
ours	15.41	2

optimizer and noam learning rate scheduler with warm-up iterations of 1302 and factor of 0.1604 are used for training MFT. We pre-train MFT with the mask and crop strategies with $M_a = 0.35$ and $C_a = 0.15$ for 100 epochs on the ICPR2022_MSR train set and augmented set, then finetune for 8 epochs with $M_a = 0.05$ and $C_a = 0.0$.

4.3. Results on ICPR2022_MSR dataset

Table 1 shows a comparison of single modality methods and our proposed WL-MSR on the ICPR2022 MSR validation set. Note that we also use OCR Filter for the OCR method to filter out noisy items. Obviously, benefiting from multimodal fusion, the proposed WL-MSR significantly outperforms the single modality methods. Compared with the ASR-only method and the OCR-only method, the CER is reduced by 13.49% and 6.20%, respectively. Final results compared with other teams on the test set of ICPR2022_MSR from `https://codalab.lisn.upsaclay.fr/competitions/2418#results` are shown in Table 2. Our method achieves an excellent CER of 15.41% and wins 2-nd place in the Multimodal Subtitle Recognition Challenge on ICPR 2022.

4.4. Ablation Study

Experimental results of ablation study could be found in Table 3. It should be noted that our baseline model uses FCENet [4] as the text detector. Improvements could be observed after implementing the proposed mask and crop strategies. Besides, features of both ASR and OCR are proved to be useful for multimodal fusion. Since subtitles are usually in rectangular boxes, FasterRCNN [3] based OCR is better than FCENet based OCR. After adding more training data by VAD augmentation, WL-MSR further achieves the CER of 19.31% on the ICPR2022_MSR validation set.

Table 3. Ablation study of WL-MSR.

Method	CER
Transformer baseline	34.13
+ mask strategy	30.35 (-3.78)
+ crop strategy	30.23 (-0.12)
+ ASR features	27.78 (-2.45)
+ OCR features	26.60 (-1.18)
+ multi-level identity embedding	26.53 (-0.07)
+ FasterRCNN detector	23.01 (-3.52)
+ VAD-based data augmentation	**19.31** (-3.70)

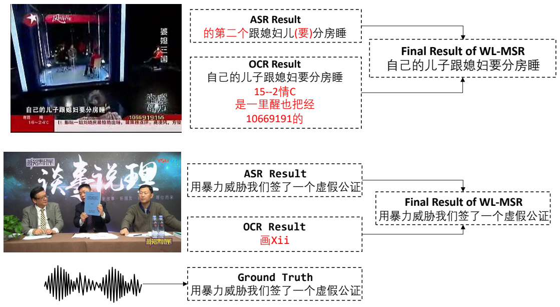

Fig. 5. An recognized example of WL-MSR.

4.5. Example

A recognized example of WL-MSR could be found in Fig. 5. The red font indicates the characters that the model recognizes incorrectly. It could be found that wrong characters may be obtained by the ASR model and there may be many noise items or even no correct subtitle in OCR results, while WL-MSR could fuse and correct the results of the two modalities.

5. CONCLUSIONS

In this paper, we focus on the task of multimodal subtitle recognition for video. To efficiently fuse the information provided by ASR and OCR models, we propose a watch-and-listen for multimodal subtitle recognition framework. We prefilter out the noisy items in the OCR results using an ASR and confidence based OCR filter. Then a multimodal fusion transformer is used to model cross-modal associations with the results and features of ASR and OCR models as input. Mask and crop strategies are adopted to release the dependence of the model on a single modality. With these techniques, our proposed framework significantly outperforms single modality methods on the ICPR2022_MSR dataset. However, the model still relies on the pre-trained ASR and OCR models instead of end-to-end, which will be studied in the future.

6. ACKNOWLEDGEMENT

This work was supported by the National Key Research and Development Program of China (No. 2020AAA0106400), National Natural Science Foundation of China (U21B2043, 62102419) and CAAI-Huawei MindSpore Open Fund.

7. REFERENCES

[1] Linhao Dong and Bo Xu, “Cif: Continuous integrate-and-fire for end-to-end speech recognition,” in *ICASSP*, 2020.

[2] Zhuoyuan Yao, Di Wu, Xiong Wang, Binbin Zhang, Fan Yu, Chao Yang, Zhendong Peng, Xiaoyu Chen, Lei Xie, and Xin Lei, “Wenet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit,” *arXiv preprint arXiv:2102.01547*, 2021.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *NIPS*, 2015.

[4] Yiqin Zhu, Jianyong Chen, Lingyu Liang, Zhanghui Kuang, Lianwen Jin, and Wayne Zhang, “Fourier contour embedding for arbitrary-shaped text detection,” in *CVPR*, 2021.

[5] Baoguang Shi, Xiang Bai, and Cong Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” *IEEE T-PAMI*, 2016.

[6] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang, “Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition,” 2021.

[7] Marvin Lavechin, Marie-Philippe Gill, Ruben Bousbib, Hervé Bredin, and Leibny Paola Garcia-Perera, “End-to-end domain-adversarial voice activity detection,” in *ICASSP*, 2020.

[8] Aditya Ramani, Asmita Rao, V Vidya, and VR Badri Prasad, “Automatic subtitle generation for videos,” in *International Conference on Advanced Computing and Communication Systems*, 2020.

[9] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *arXiv preprint arXiv:2006.11477*, 2020.

[10] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao, “Detecting text in natural image with connectionist text proposal network,” in *ECCV*, 2016.

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *ICML*, 2006.

[12] Zbigniew Wojna, Alexander N Gorban, Dar-Shyang Lee, Kevin Murphy, Qian Yu, Yeqing Li, and Julian Ibarz, “Attention-based extraction of structured information from street view imagery,” in *ICDAR*, 2017.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, “Attention is all you need,” *NIPS*, 2017.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” in *NACACL-HLT*, 2019.

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, “UNITER: universal image-text representation learning,” in *ECCV*, 2020.

[16] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao, “Vinvl: Revisiting visual representations in vision-language models,” in *CVPR*, 2021.

[17] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, “pyannote.audio: neural building blocks for speaker diarization,” in *ICASSP*, 2020.

[18] S. Huang et al., “Icpr 2022 challenge on multi-modal subtitle recognition,” in *ICPR*, 2022, pp. 4974–4980.

[19] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, “fairseq: A fast, extensible toolkit for sequence modeling,” in *NACACL-HLT*, 2019.

[20] Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng, “AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline,” in *O-COCOSDA*, 2017.

[21] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al., “Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019,” in *ICDAR*, 2019.

[22] Tai-Ling Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, and Shi-Min Hu, “Chinese text in the wild,” *arXiv preprint arXiv:1803.00085*, 2018.

[23] Jonathan T Barron, “Continuously differentiable exponential linear units,” *arXiv preprint arXiv:1704.07483*, 2017.