Multimodal genre recognition of Chinese operas with hybrid fusion

Fan, Tao
fantao0916@gmail.com
Nanjing University, China, People's Republic of; University of Bern, Switzerland

Wang, Hao
ywhaowang@nju.edu.cn
Nanjing University, China, People's Republic of

Hodel, Tobias
tobias.hodel@unibe.ch
University of Bern, Switzerland

Introduction

Chinese operas are the treasure of Chinese culture, and also the important component of world intangible cultural heritage (Zhang et al. 2008). In the contemporary China, listening operas in music platforms is a very popular way among the young and old people. Due to the support of the government and the public's attention, Chinese opera is in a new stage with high development. The creative enthusiasm is tremendously motivated, and more and more works of Chinese opera with different genres are produced. However, it is very hard for the public without the rich knowledge of opera to recognize the genre correctly. Assuming the scenario where people hear the opera from the music platform, however, the platform only provides the name without the information of the genre, which is not beneficial for the listeners who want to have the deep experience of the genre that the work belongs to. Hence, utilizing an efficient model to recognize their genres automatically is extremely necessary and important, which is beneficial for the broadcast and sustainable development of Chinese operas.

Currently, there have been studies exploring the automatic recognition of operas. Related researches can be divided into unimodal-based (lyrics, audios) and multimodal-based methods (Jin et al. 2022; Luo 2021). If we only take audios or lyrics-based methods to recognize the genre, the auxiliary and complementary semantics information will be ignored. Furthermore, if we take a multimodal way, the information in lyrics and audios can be utilized simultaneously but it requires us to provide the complete multimodal information. Usually, lyrics of operas are provided by the labour most of the time which guarantees the quality and precision. However, in real scenarios, the situation that lyrics of operas are missing is common. If we utilize the common-used API to transform the audios into lyrics automatically, the performance is extremely bad and meanings of operas will be changed completely due to the unique singing characteristics and arias of Chinese operas.

Therefore, we wonder whether it is possible to build a unified model to recognize the genre of Chinese operas with multimodal information and tackle the problem of missing modality.

The proposed model

Based on that, we build a unified **M**ultimodal **G**enre **R**ecognition of Chinese operas with **H**ybrid **F**usion (**MGRHF**), shown in Figure 1, which can utilize the multimodal information effectively learned from lyrics and audios in Chinese operas, and also fix the missing modality problem. The model structure consists of the lyric-based, audio-based, lyric-audio-based model, and hybrid fusion technique, which is composed of the intermediate-level and decision-level fusion. In the lyric-based model, we first use Bidirectional Encoder Representation Transformers (BERT) (Devlin et al. 2019) to encode lyrics of operas, to obtain the representative features. Then we feed the representative features into the context learner, Bidirectional Long Short-Term Memory networks (BiLSTM) (Hochreiter / Schmidhuber 1997), to excavate the contextual and semantics features in lyrics. Lastly, the features with rich contextual textual features are input to layers composed of a fully connected (FC) layer with *relu* activation and *softmax* layer. In the audio-based model, deep spectrum features extracted from spectrograms of audios, through the ResNet model (He et al. 2016) pretrained on ImageNet, are employed as the representation of opera audios. Next, the spectrum features are fed into two FC layers with different activation functions (*relu* and *softmax*), to predict the genre of operas. The lyric-audio-based model utilizes the networks learning contextual semantics and visual features, and an intermediate fusion strategy is employed to concatenate features from lyrics and audios. Also, two FC layers (with *relu* and *softmax*) are added to perform non-linear activation and give the prediction probabilities. Lastly, we apply the decision-level fusion technique to obtain the final genre prediction result. Specifically, we denote the prediction probabilities on a Chinese opera from the lyric-, audio-, and lyric-audio-based models as p l , p a , and p m . The final genre label l g e n r e is predicted by the following equation:

$$l_{genre} = \begin{cases} argmax(\alpha p_l + \beta p_a + \gamma p_m), & input: (audio, lyric) \\ argmax(p_a), & input: audio \\ argmax(p_l), & input: lyric \end{cases}$$

α , β , and γ are the optimal weight parameters through the grid-searching method, ranging from 0 to 1, the searching step is 0.1, and the sum of all weights is 1. If one modality is missing, we just need to set the corresponding weight and multimodality weight to 0, which can help us to solve the missing modality problem.

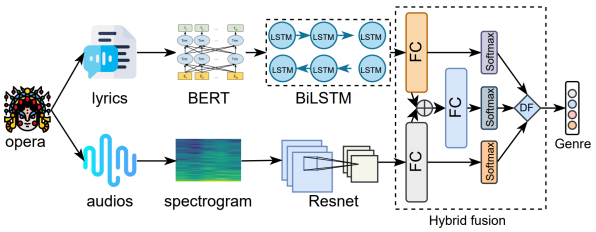

Figure 1. The structure of MGRHF

Dataset of Chinese operas

In this paper, four nation-level Chinese operas genres are selected, which are *Beijing opera*, *Kun opera*, *Yue opera*, and *Jin Opera*. The four operas are the famous and popular operas in China, and can be seen as the representatives of Chinese operas. NetEase platform is chosen as the data source to scrape audios and

lyrics of four operas. Then, we filter the data and segment the audios based on the sentences in the lyrics into several clips. Lastly, the number distribution of four operas is illustrated in the Table 1.

Table 1. The number distribution of different operas

Dataset	*Beijing Opera*	*Kun opera*	*Yue opera*	*Jin opera*
Number	1,027	912	969	870

Experiments

Here, we select classic BERT, ResNet, EF-LSTM, MGR, and CNN model as baseline models. Experimental results are shown in Table 2. It can be seen that MGRHF obtains the superior results compared with baseline models considering all evaluated metrics, which shows the effectiveness of our proposed model. In addition, in the comparison with MGR, we find that the performance of MGRHR utilizing the proposed hybrid technique is better, which demonstrates its superiority in the genre recognition of Chinese operas. Also, it can be observed that multimodal fusion models have an advantage over unimodal models.

Table 2. Experimental results of MGRHF and baseline models. EF-LSTM concatenates the textual and spectrum features, which are input to LSTM model. MGR is the model without decision-level fusion.

Models	*Accuracy* (%)	*Precision* (%)	*Recall* (%)	*F1* (%)
BERT	90.39	90.71	91.09	90.86
ResNet	90.00	89.45	89.07	89.08
EF-LSTM	96.54	96.03	96.50	96.22
MGR	97.31	97.34	97.00	97.13
CNN	89.62	90.04	90.38	89.99
MGRHF	**97.69**	**97.85**	97.41	97.59

Conclusion

In summary, we present our model MGRHF for the automatic genre recognition of Chinese operas, and conduct the empirical experiments. Experimental results demonstrate the strong performance of MGRHF. In the future, case studies will be performed and MGRHF will be tested in a wider dataset.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 72074108), Special Project of Nanjing University Liberal Arts Youth Interdisciplinary Team (010814370113), Jiangsu Young Social Science Talents, “Tang Scholar” of Nanjing University, and China Scholarship Council (202206190149).

Bibliography

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* , 4171–4186. https://doi.org/10.18653/v1/N19-1423

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 770–778. https://doi.org/10.1109/CVPR.2016.90

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation* , *9* (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Jin, C., Song, Z., Xu, J., & Gao, H. (2022). Attention-Based Bi-DLSTM for Sentiment Analysis of Beijing Opera Lyrics. *Wireless Communications and Mobile Computing* , *2022* , e1167462. https://doi.org/10.1155/2022/1167462

Luo, W. (2021). Analysis of Artistic Modeling of Opera Stage Clothing Based on Big Data Clustering Algorithm. *Security and Communication Networks* , *2021* , e5349916. https://doi.org/10.1155/2021/5349916

Zhang, Y.-B., Zhou, J., & Wang, X. (2008). A study on Chinese traditional opera. *2008 International Conference on Machine Learning and Cybernetics* , *5* , 2476–2480. https://doi.org/10.1109/ICMLC.2008.4620824