Easy Travelogue: A Travelogue Editor with Automatic Image Recommendation and Insertion

Fan Yu
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
yf@smail.nju.edu.cn

Huanyu Xing
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
xosmos@foxmail.com

Jia Bei*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
beijia@nju.edu.cn

Tongwei Ren
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
rentw@nju.edu.cn

ABSTRACT

Travelogues are a common media form that incorporates both text and images. Typically, they are composed after the completion of a travel period. Creating a travelogue demands substantial time and effort, particularly in the curation of suitable images from the extensive collection of photos taken during the journey to complement the text. Consequently, we have developed and implemented Easy Travelogue, a travelogue editor that utilizes visual and language models. It offers real-time image suggestions while writing the text and can automatically insert fitting images into the finished content. The editor is versatile and can be readily utilized for personal travelogues, travel blogs, and various social media platforms, facilitating users in effortlessly sharing and showcasing their travel experiences.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Travelogue editing, image recommendation, text-image retrieval, vision language model

ACM Reference Format:
Fan Yu, Huanyu Xing, Jia Bei, and Tongwei Ren. 2023. Easy Travelogue: A Travelogue Editor with Automatic Image Recommendation and Insertion. In *ACM Multimedia Asia 2023 (MMAsia '23), December 6–8, 2023, Tainan, Taiwan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3595916.3626353

1 INTRODUCTION

With the development of the tourism industry, there is an increasing demand for travel information [3, 4, 6]. People are also willing to share their travel experience on the Internet. Travelers usually capture a considerable number of real-time photos as a means of preserving memories during their journeys, yet they do not engage in real-time travelogue writing. Generally, they organize their travel experiences after the trip, by which time they have accumulated a large number of travel images. The process of writing a travelogue involves selecting appropriate images from a large pool of photos and inserting them into the text, which requires a significant amount of time and effort. Existing intelligent travelogue editors typically select images based on metadata such as time and location of the photos, but they lack analysis of the semantic content of the images. Therefore, a significant amount of manual secondary screening is still required. However, with the development of vision language models (VLMs), the semantic space of images and text can be aligned, providing technical support for automatic image matching to travelogues. A travelogue editor that can automatically select and recommend images based on the semantic content of both images and text can greatly reduce the cost of travelogue producing. Considering the inconvenience of typing texts and selecting images on mobile devices, it is necessary to implement automatic image insertion for long-form content. Thus, as the Figure 1 shows, we design and implement a travelogue editor named Easy Travelogue with automatic image recommendation and insertion based on VLMs.

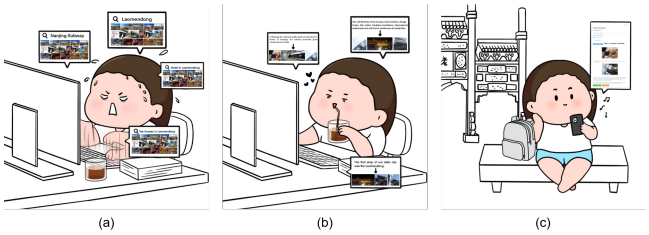

Figure 1: Common editor *vs.* Easy Travelogue. (a) Searching images for editing travelogue with common editor. (b) Real-time image recommendation with Easy Travelogue. (c) Off-line image insertion with Easy Travelogue.

*Corresponding author.

MMAsia '23, December 6–8, 2023, Tainan, Taiwan

ACM ISBN 979-8-4007-0205-1/23/12.
https://doi.org/10.1145/3595916.3626353

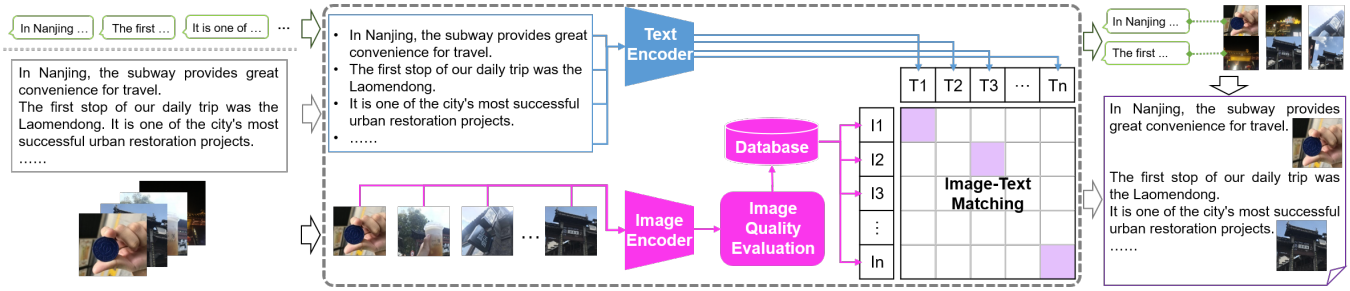

Figure 2: Framework for matching of travelogue text and images.

2 AUTOMATIC IMAGE RECOMMENDATION AND INSERTION

As shown in Figure 2, the key technology in our travelogue editor is the matching of travelogue text and images.

Vision and language alignment. Recently, models like CLIP [2] and ALIGN [1] trained on large-scale datasets learn universal cross-modal representations. Such models leverage a dual-encoder architecture to align visual and language representations of image and text pairs with contrastive learning. In our editor, we exploit the Chinese-CLIP [5], which is finetuned on Chinese corpus. Since the text and images in the real travelogue are not strictly matched according to a fixed set of rules, we do not finetune the Chinese-CLIP with travelogue corpus.

Image feature indexing. Given the potential abundance of images as recommendation and insertion candidates, we promptly extract image features using Chinese-CLIP upon importing the images and save the index in a database. Using the extracted image features, we calculate similarities between images and assess the quality of each image. Consequently, we recommend only one image with the highest quality score out of all similar images, enhancing the effectiveness of image recommendation.

Image-text matching. In order to match images with text, we calculate the cosine similarities between images and text and select appropriate images for the travelogue. In the case of image recommendation, images are selected for each sentence. However, in the case of image insertion, the most suitable sentence is selected for each image or a group of similar images.

3 EASY TRAVELOGUE

Easy Travelogue provides real-time image recommendation for travelogue text and automatic images insertion for off-line text. Users can create material repositories to store text and images. On the basis of the material repository, users can generate travelogues in real-time mode or off-line mode. Easy Travelogue supports both Chinese and English.

3.1 Off-line image insertion

Considering the convenience of mobile device usage, we provide off-line image insertion into text (Figure 3). Users can verbally share their travel experience, and the editor will save the text. After users select text and image materials, the editor filters out the desired travel images for insertion and finally creates a comprehensive travelogue.

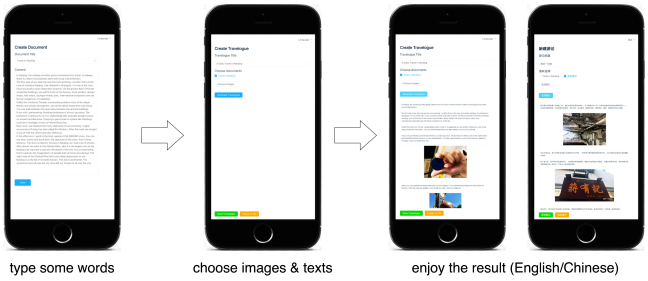

Figure 3: Screenshot of off-line insertion of images that match the edited text.

Figure 4: Screenshot of real-time recommendations of images that match the edited text.

3.2 Real-time image recommendation

Users can input text online in the text box, and the editor will recommend the most relevant images based on the content of the current input sentence, presenting the images to the user in descending order of visual-textual similarity (Figure 4). Users can select one or more images for insertion. Moreover, the editor automatically group images with high similarities into the same image group. Once an image is inserted, the image will no longer be recommended in the subsequent editing process.

4 CONCLUSION

In this paper, we designed and implemented a travelogue editor named Easy Travelogue with automatic image recommendation and insertion. This system supports real-time image recommendation for edited text. Also the system can generate complete travelogues by inserting images automatically into text translated from speech.

ACKNOWLEDGMENTS

This work is supported by National Natural Science Foundation of China (62072232), the Fundamental Research Funds for the Central Universities (021714380026) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[3] Xi Shao, Guijin Tang, and Bing-Kun Bao. 2019. Personalized travel recommendation based on sentiment-aware multimodal topic model. *IEEE Access* 7 (2019), 113043–113052.

[4] Junyi Wang, Bing-Kun Bao, and Changsheng Xu. 2019. Sentiment-aware multimodal recommendation on tourist attractions. In *MultiMedia Modeling: 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8–11, 2019, Proceedings, Part I 25*. Springer, 3–16.

[5] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive vision-language pretraining in Chinese. *arXiv preprint arXiv:2211.01335* (2022).

[6] Xin Zhang, Xiaoqian Lu, Xiaolan Zhou, and Chaohai Shen. 2022. Reconsidering tourism destination images by exploring similarities between travelogue texts and photographs. *ISPRS International Journal of Geo-Information* 11, 11 (2022), 553.