Exploring User Preferences for Seamless Scene Text Translation in Video

ANONYMOUS AUTHOR(S)

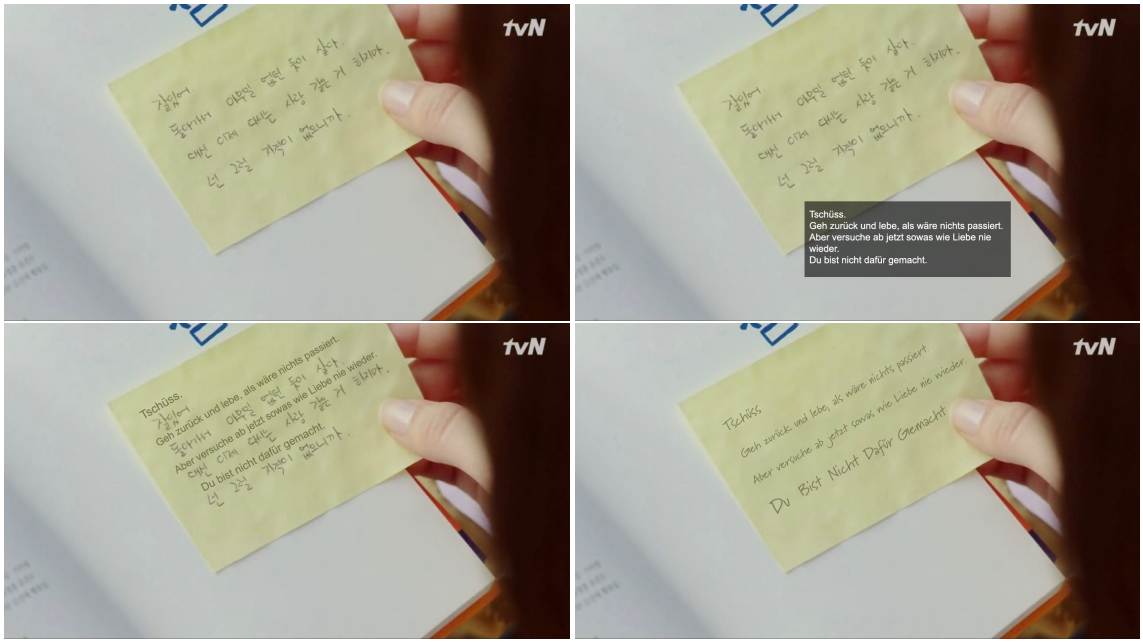

Fig. 1. (1) Original scene of a long handwritten text; (2) Display style *Overlay*; (3) Display style *Integration*; (4) Display style *Replacement*.

Our research explores different methods for embedding scene text translations in video content, particularly the potential of AI-driven solutions that preserve aesthetic consistency through seamless text replacement, such as of handwritten notes or street signs. In a user study with 24 participants, we compared three approaches for displaying scene text translations against the original footage. Our findings reveal highly variable user preferences, influenced by both individual differences and the contextual significance of the scene text. These results highlight the need for customizable solutions and context-aware translation strategies in future developments.

CCS Concepts: • **Human-centered computing** → **Empirical studies in visualization**; **Visualization theory, concepts and paradigms**; **Accessibility systems and tools**; • **Information systems** → **Multimedia and multimodal retrieval**; • **Applied computing** → **Media arts**; • **Computing methodologies** → **Computer vision tasks**; **Natural language processing**.

Additional Key Words and Phrases: Scene Text Translation, Video Text Editing, Optical Character Recognition (OCR), Multilingual Accessibility, Text Integration, Film, Video, Computer Vision

Manuscript submitted to ACM

ACM Reference Format:
Anonymous Author(s). 2018. Exploring User Preferences for Seamless Scene Text Translation in Video. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Subtitles enhance the accessibility of video content for non-native speakers and the deaf or hard of hearing while improving clarity also for those who understand the language. Advancements in Artificial Intelligence (AI) have enabled the automation of subtitle generation, making it possible to easily access movies, TV shows and short-form video content in foreign languages. However, conventional subtitles typically focus on spoken dialogue, often neglecting scene text that is crucial for narrative comprehension. This issue becomes particularly evident in scenes featuring written communication, such as text messages or letters, that characters react to without verbalizing the content.
A potential solution involves OCR-based text detection, followed by translation and generative AI-powered text replacement, ensuring that the original typographic styling is preserved. This approach would allow seamless translation of scene text while maintaining aesthetic continuity, including font, size, color, arrangement, and perspective. Current scene text editing technologies are, however, designed for images and have yet to be fully adapted for video. In our research, we aim to evaluate whether scene text editing for video could enhance the viewing experience of foreign videos.

2 BACKGROUND & RELATED WORK

2.1 Text in Video

Text in video can generally be classified into two categories [8, 10]:

- **Superimposed**, or **graphic text** encompasses digitally added elements during post-production, including titles, subtitles, credits and informational overlays.
- **Scene text** appears naturally within recorded environments, including street signs, letters, license plates, and product labels.

While superimposed text in videos can be manually adapted for different languages and is easier to detect and replace automatically, modifying scene text in post-production poses greater challenges. Variations in perspective, lighting conditions, and complex visual backgrounds make accurate detection and seamless replacement significantly more difficult.

2.2 Optical Character Recognition

Optical Character Recognition (OCR) extracts handwritten and typed text from various sources like scanned documents, PDFs, or images for editing and data retrieval. OCR methods have evolved from traditional algorithms to machine learning-based approaches, such as *Tesseract*, a widely used open-source OCR engine, and cloud-based solutions like *Google Vision* [1, 7]. Other tools, such as Amazon Rekognition Video[1], offer similar functionalities, with the latter specializing in real-time analysis within the AWS ecosystem. Slik et al. [11] introduced a real-time system to recognize 'burned-in' subtitles and address background interference. Later research tackled complex backgrounds [12] and

[1]https://aws.amazon.com/de/rekognition/video-features/

multimodal recognition [6]. OCR also facilitates scene text translation, as seen in Google Translate's real-time overlay feature and Yang et al.'s [13] prototype for Chinese sign translation.

2.3 Text Editing in Images

Recent progress in generative image modeling and style transfer has greatly enhanced the ability to edit text in images while preserving background context, font style, and color. Scene text editing models such as SRNet [9] and SWAPText [14] aimed to retain text style while modifying content but required target style images for training. More recent handwriting-focused models [4, 5] have explored self-supervised and few-shot learning approaches. However, they often struggle with domain variability and fail to generalize effectively across both printed and handwritten text. TextStyleBrush [9], a research project by Meta, addresses these limitations by leveraging self-supervised learning and a StyleGAN2-based architecture, enabling one-shot style transfer while maintaining text structure and readability across diverse domains. As diffusion models have become the dominant approach for image generation, new diffusion-based technologies have emerged for generating text in images [2, 3], further expanding the possibilities for complex text editing in images.

3 ON-SCREEN SCENE TEXT TRANSLATION

Our research explores methods for implementing multilingual support for scene text in videos. We compare three approaches to scene text translation (for examples of each type of translation, see Figure 1).

- **Overlay**: This method resembles traditional subtitles, where the translated text is superimposed at the bottom of the screen. To improve readability, a semi-transparent contrasting background is applied.
- **Integration**: The translation appears directly below the original text within the scene. While the font remains a standard typed style, as in the Overlay method, its color, position, and size are adjusted to blend better with the visual context. This approach maintains the visibility of the original text while achieving a more natural integration.
- **Replacement**: This method involves direct image manipulation, replacing the original text with the translated text while preserving its visual style. The color, placement, font, and size are carefully adjusted to match the original appearance, making the modification barely noticeable.

4 USER STUDY

In a preliminary user study, we explored the potential of seamless text replacement in video. To assess user demand for this technology, we conducted a study with 24 participants, evaluating different variants of film scenes that were manually edited by us. Each variant is based on a different presentation style - Overlay, Integration and Replacement. The users then rated the original and the three translated variants. This provided insights into the most favored style for displaying scene text translations.

5 PRELIMINARY RESULTS AND INSIGHTS

In a scene featuring a handwritten note (see Figure 1), the Original version (M = 1.58, SD = 0.83) was the least favored. Interestingly, the three translation styles received similar mean scores, with no single method emerging as the clear favorite. Overlay achieved the highest mean rating (M = 3.71, SD = 1.2), followed closely by Replacement (M = 3.54, SD = 1.25) and Integration (M = 3.38, SD = 1.28). When asked which option they preferred, 10 participants selected

Overlay. Replacement was nearly as popular, receiving 9 votes. 5 participants selected the Integration option as their top choice, and none selected the Original version. The participants articulated their discontent with the initial version, citing the language as a significant hindrance due to its inaccessibility, which resulted in the perception that crucial information was not readily available. Even when the content was not essential to the plot, they still wanted the option to access it. Only one person rated this version highly, appreciating its "authenticity" and noting that handwriting adds personality and context to a scene. The Overlay method was praised for its readability without significantly altering the original image. Participants described it as the "quickest and most pleasant" way to read, as it preserved the viewing experience while ensuring a clear separation between the original text and its translation. Some valued how it retained the "charm of an international film," while others found it less immersive and somewhat visually intrusive. Integration was appreciated for its natural and immersive look, as well as its usefulness for language learning by allowing a direct comparison between the original and translated text. However, it was often criticized for being difficult to read and visually cluttered. The Replacement option was favored for its readability and seamless integration into the scene. Some appreciated how it minimized distractions, particularly in fast-paced sequences. However, others disliked the removal of the original text, arguing that it reduced authenticity and made nuanced translations harder to convey—especially in languages that require contextual explanations. A film industry professional also pointed out that replacing text compromises the ability to differentiate between original footage and altered elements. While some considered this the "coolest" option, many had a preference for a version that allowed the original text to be revealed when needed.

However, our study also revealed that preference for translation types varies depending on the type of scene text. For instance, when it came to street or shop signs, participants who initially favored Overlay stated that they would switch to either Original or Integration. This shift stems from the lower narrative importance of such text compared to plot-critical elements, as well as concerns about visual clutter and potential confusion if too many signs were translated. For text added in post-production, such as superimposed chat bubbles, most participants preferred Replacement, primarily because it allows for faster reading and a more seamless integration into the scene.

6 OUTLOOK

Advancements in OCR, natural language processing and AI-based text editing for video will soon enable the automated translation of scene text, making movies and other video content accessible in multiple languages. However, our study shows that preferences for how translations are embedded vary among participants and shift depending on the context of the scene text. Therefore, AI models could be leveraged to determine the most suitable translation method for each scenario, and customization options need to be incorporated into the design to accommodate user preferences.

REFERENCES

[1] 2023. OCR with Google Vision API and Tesseract. *The Programming Historian* 12 (2023). https://doi.org/10.46430/phen0109

[2] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2024. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*. Springer, 386–402.

[3] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2024. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems* 36 (2024).

[4] Brian Davis, Chris Tensmeyer, Brian Price, Curtis Wigington, Bryan Morse, and Rajiv Jain. 2020. Text and style conditioned GAN for generation of offline handwriting lines. *arXiv preprint arXiv:2009.00678* (2020).

[5] Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. 2020. Scrabblegan: Semi-supervised varying length handwritten text generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4324–4333.

[6] Shan Huang, Shen Huang, Li Lu, Pengfei Hu, Lijuan Wang, Xiang Wang, Jian Kang, Weida Liang, Lianwen Jin, Yuliang Liu, Yaqiang Wu, and Yong Liu. 2022. ICPR 2022 Challenge on Multi-Modal Subtitle Recognition. In *2022 26th International Conference on Pattern Recognition (ICPR)*. 4974–4980. https://doi.org/10.1109/ICPR56361.2022.9956308

[7] Geeta S Hukkeri, R H Goudar, Prashant Janagond, and Pooja S Patil. 2022. Machine Learning in OCR Technology: Performance Analysis of Different OCR Methods for Slide-to-Text Conversion in Lecture Videos. *International Journal of Advanced Computer Science and Applications* 13, 8 (2022). https://doi.org/10.14569/IJACSA.2022.0130839

[8] Keechul Jung, Kwang In Kim, and Anil K Jain. 2004. Text information extraction in images and video: a survey. *Pattern recognition* 37, 5 (2004), 977–997.

[9] Praveen Krishnan, Rama Kovvuri, Guan Pang, Boris Vassilev, and Tal Hassner. 2023. Textstylebrush: transfer of text aesthetics from a single example. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 7 (2023), 9122–9134.

[10] P Shivakumara, N Vinay Kumar, DS Guru, and Chew Lim Tan. 2014. Separation of graphics (superimposed) and scene text in video frames. In *2014 11th IAPR International Workshop on Document Analysis Systems*. IEEE, 344–348.

[11] MARCO Slik, HANS Jongebloed, and Mark Van Staalduinen. 2013. Video based OCR: A case study of real time in-screen subtitle recognition. *Tech. Rev* (2013).

[12] Hongyu Yan and Xin Xu. 2020. End-to-end video subtitle recognition via a deep Residual Neural Network. *Pattern Recognition Letters* 131 (March 2020), 368–375. https://doi.org/10.1016/j.patrec.2020.01.019

[13] Jie Yang, Xilin Chen, Jing Zhang, Ying Zhang, and Alex Waibel. 2002. Automatic detection and translation of text from natural scenes. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 2. II–2101–II–2104. https://doi.org/10.1109/ICASSP.2002.5745049

[14] Qiangpeng Yang, Jun Huang, and Wei Lin. 2020. Swaptext: Image based texts transfer in scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14700–14709.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009