부산대학교

정보컴퓨터공학부

자연스러운 번역 텍스트 합성 착수보고서

이동훈, 이승재, 문경환

Supervisor: 전상률

May 14, 2025

Contents

Chapter 1

Introduction

1.1 Research Background

전 세계적으로 디지털 콘텐츠의 수요가 급증함에 따라, 영상 내 텍스트 번역 기술의 중요성이 더욱 부각되고 있다. 특히 COVID-19 팬데믹 이후 비대면 콘텐츠 소비가 일상화되면서, 다양한 언어권 사용자들이 다국어 기반의 영상 콘텐츠를 접하게 되었고, 이에 따라 보다 정교하고 자연스러운 번역 기술에 대한 필요성이 제기되고 있다. 기존의 번역 시스템은 이미지 내 문자를 인식(OCR)한 후, 해당 문장을 번역하여 별도의 텍스트 박스에 삽입하는 방식이 일반적이었다. 이와 같은 방식은 기계적 번역과 정보 전달에는 유효할 수 있으나, 시각적 자연스러움과 맥락적 조화 측면에서는 한계가 명확하다. 이러한 전통적인 방식은 텍스트의 색상, 위치, 배경 질감, 광원 정보 등을 반영하지 못하며, 번역된 텍스트가 원본 이미지로부터 시각적으로 분리되어 보이는 문제를 야기한다. 그 결과, 사용자로 하여금 이질적인 인상을 유발하며, 콘텐츠에 대한 몰입도 및 정보 수용력을 저하시킬 수 있다. 예컨대, 제품 패키지나 간판에 포함된 외국어 문장을 번역할 경우, 단순한 오버레이 방식은 원본 디자인을 훼손하거나 정보 전달의 직관성을 저해하는 결과로 이어진다.

이러한 한계를 극복하기 위해 최근에는 번역된 텍스트를 원본 이미지의 시각적 속성과 정합되도록 자연스럽게 합성하는 기술이 주목받고 있다. 본 기술은 단순히 문자 정보를 변환하는 수준을 넘어, 배경 이미지의 질감, 조도, 왜곡, 타이포그래피 스타일 등을 고려하여 텍스트를 재구성함으로써, 영상 콘텐츠의 시각적 일관성과 감성적 품질을 동시에 제고할 수 있는 가능성을 내포한다. 특히 방송, 전자상거래, 관광, AR/VR 등 다양한 응용 분야에서 이와 같은 자연스러운 번역 텍스트 합성 기술은 사용자의 몰입 경험을 향상시키고, 언어 장벽을 효과적으로 해소할 수 있는 핵심 요소로 간주되고 있다.

1.2 Research Purpose

본 연구에서는 기존의 기계 번역 및 OCR 기반 텍스트 삽입 방식이 지니는 한계를 분석하고, 자연스러운 번역 텍스트 합성을 위한 기술적 요구 사항과 구현 방식에 대해 연구하고자 한다. 나아가 이를 활용한 어플리케이션 서비스를 제안하고 구현하여 사용자 입력 이미지에 따른 자연스러운 번역 텍스트 합성이 가능하도록 하는 것이 최종 목표이다.

Chapter 2

Problem Definition and Requirements

2.1 Problem Statement

본 연구의 목적은 사용자 입력 이미지에 따른 자연스러운 번역 텍스트 합성이 가능하도록 하는 것이다. 이미지는 사용자로부터 입력받으며, 입력받은 이미지의 텍스트에 따른 번역 텍스트 재합성을 목표한다.

2.2 Functional Requirements

사용자는 동작에 적합한 이미지를 입력으로 넣을 수 있다. 본 프로그램은 입력에 따라 다음과 같은 작업을 수행한다.

- 입력받은 이미지의 텍스트를 번역.
- 이미지의 텍스트를 번역해 재합성.

입력받는 이미지는 다양한 크기가 가능하여야 하며, 최종 출력은 번역된 텍스트를 제외하고 원본 이미지의 형태를 유지할 수 있어야 한다.

2.3 Non-functional Requirements

시스템의 출력 결과는 이미지 파일로 저장되어야 하며, 이에 따라 모델의 프로세스에서 출력된 모든 결과는 csv 혹은 이미지 파일로 저장할 수 있어야 한다.

2.4 Input/Output Specifications (Data or Interaction)

본 모델은 인식, 번역, 제거, 이전 및 합성 단계로 구성되며, 이에 따른 입출력은 표 2.1과 같다.

Models	**Description**	**Input**	**Output**
Translate	번역	Segmented image	Translated text
Remove	글자 삭제	Image	Image
Font	원본 폰트와 생성한 **폰트** 형태, 굵기, 크기, 정렬 비교	List of images	크기(float), 굵기(int), 폰트(category)
Synthesis1	오브젝트의 특성 합성	Segmented image	Image
Synthesis2	글자 합성	Image, text image	Image

Table 2.1: Agent 기능별 입력 및 출력 정의

Chapter 3

System Design and Architecture

3.1 Overall System Architecture

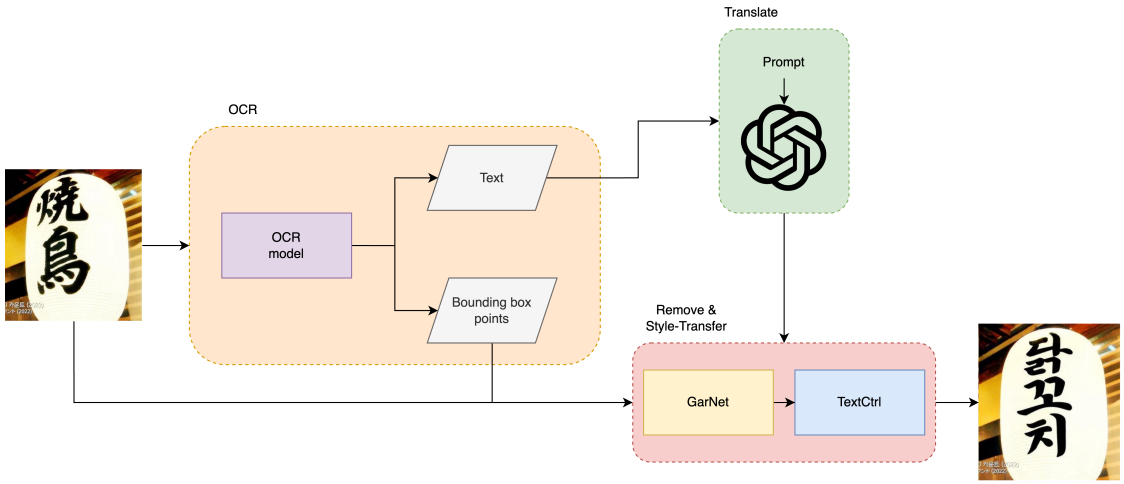

Figure 3.1: 모델 아키텍쳐 개요

프레임워크는 총 3단계로 구성되며, 텍스트 추출, 번역, 삭제 및 합성 단계로 구성한다. 시스템의 전반적인 구조는 그림 3.1과 같다.

3.2 Simulation Algorithm Design and Tools

추출 단계에서는 OCR 모델을 사용해 텍스트와 텍스트의 위치를 뽑아낸다. OCR 모델은 PaddleOCR을 사용한다. PaddleOCR는 Baidu의 딥러닝 프레임워크인 PaddlePaddle을 기반으로 개발된 오픈소스 광학문자인식(OCR) 시스템으로, 다양한 언어와 장치에서의 텍스트 검출 및 인식을 지원한다. 이 모델은 텍스트 검출, 방향 분류, 텍스트 인식의 세 가지 주요 단계로 구성되며, 각 모듈이 독립적으로 동작하도록 설계되어 효율적인 개선과 확장이 가능하다. PaddleOCR의 PP-OCR 시리즈는 경량화와 정확도 향상을 동시에 추구하며, Collaborative Mutual Learning, CopyPaste 증강, LCNet 등의 기술을 통해 성능을 지속적으로 향상시켰다. 특히, PP-OCRv3에서는 LK-PAN과 RSE-FPN을 통한 검출 정확도 향상, SVTR-LCNet과 TextRotNet을 통한 인식 성능 개선이 이루어져 모바일 및 임베디드 환경에서도 높은 성능을 유지하게 한다 Li et al. (n.d.). 모델을 통해 얻은 데이터는 그림 3.1과 같이 각각 번역 단계와 삭제 및 합성 단계로 포워딩된다.

번역 단계에서는 chat gpt를 활용하며, 번역 프롬프트를 통해 입력으로 받은 텍스트를 번역한다.

삭제 및 합성 단계에서는 2가지 모델이 사용된다. 삭제는 GaRNet(Gated Attention and Region of Interest Network)을 사용하며, 해당 모델은 장면 텍스트 제거를 위한 딥러닝 기반 모델로, 입력 이미지에서 텍스트를 효과적으로 제거하여 배경을 자연스럽게 복원하는 데 중점을 둔다. 모델 구조는 인코더-디코더 구조를 기반으로 하며, 인코더는 Gated Attention 모듈을 포함한 잔차 블록(residual block)으로 구성된다. Gated Attention 모듈은 두 개의 공간 주의(spatial attention) 분기를 통해 텍스트 스트로크와 그 주변 영역을 각각 탐지하고, 학습 가능한 파라미터를 통해 이 두 영역의 가중치를 조절한다 Lee and Choi (2022). 한편 합성은 TextCtrl을 사용한다. 본 모델은 텍스트의 배경, 전경, 글꼴, 색상 등의 스타일 요소를 세분화하여 분리하는 텍스트 스타일 해리 기법을 적용하여 편집된 텍스트가 원본 이미지의 스타일과 일관되도록 하며, 텍스트의 글리프 구조를 견고하게 표현하여 렌더링 정확도를 향상시킨다. 또한, 추론 과정에서 원본 이미지의 세밀한 특징을 활용함으로써 스타일 일관성을 강화한다 Zeng et al. (2024).

두 모델을 사용하여 텍스트를 삭제하고 번역된 텍스트를 폰트에 적응한 형태로 스타일 이전한 이미지를 합성하여 최종 이미지를 생성한다.

Chapter 4

Constraints and Risk Analysis

4.1 Realistic Constraints

- 여러 모델을 사용함에 따른 계산 비용 증가에 따른 학습 및 시현의 어려움.
- 인공지능 모델의 다양한 상황(e.g., 텍스트가 세로 정렬)에 대한 인식의 어려움.
- 비교 프레임워크의 부족으로 인한 성능 평가의 어려움.
- 곡률이 있는 입체 표면의 텍스트 합성의 어려움.

4.2 Countermeasures and Mitigation Strategies

- 각 단계를 나누어 한 단계에 사용되는 모델 최소화.
- 단계의 작업을 통합할 수 있는 모델 사용.
- 모델 비교를 통한 경량화된 모델 사용.
- OCR 결과를 패치로 나누어 각 어절에 변환을 사용함으로 텍스트 변환.
- 서비스: 각 단계별 검토 과정을 추가하여 성능 평가에 반영.
- 표면의 곡률을 분석할 수 있는 에이전트 도입.

Chapter 5

Implementation Schedule and Role Assignment

5.1 Development Timeline

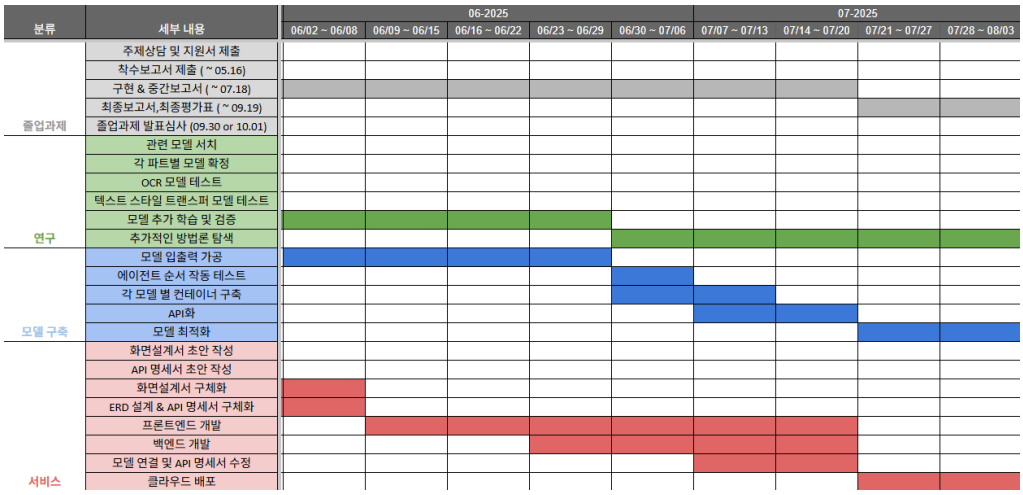

Figure 5.1: 타임라인에 따른 착수 일정 (상반기)

타임라인에 따른 착수 일정은 그림 5.1, 그림 5.2와 같다.

5.2 Team Member Roles

이동훈

- 프로젝트 UI/UX 구상
- API 명세서 작성
- 프론트/벡엔드 개발

이승재

- OCR 모델 및 텍스트 변환 모델 실험
- 모델 연결 위한 도커 이미지 정리

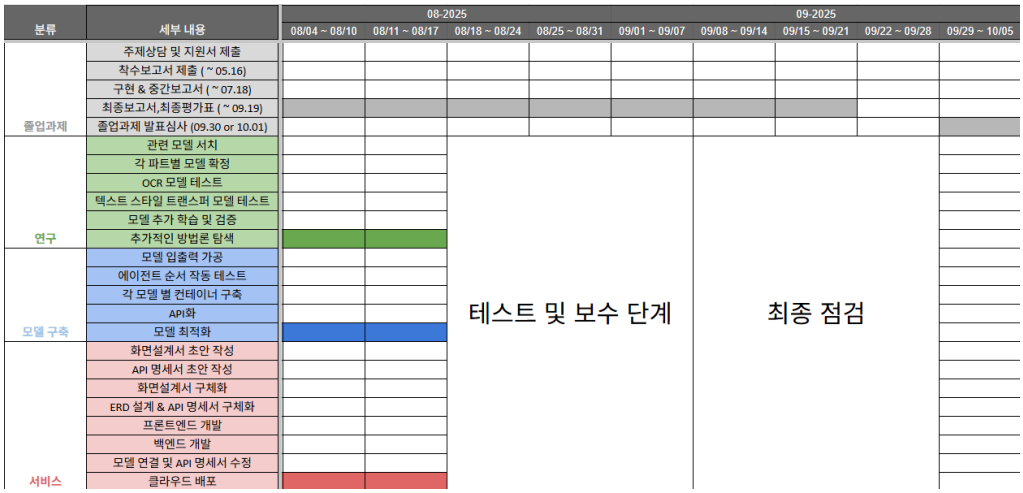

Figure 5.2: 타임라인에 따른 착수 일정 (하반기)

- 벡엔드 개발

문경환

- 실험 환경 구축 및 모델 실험
- 보고서 및 문서 작성
- 방법론 추가 탐색

References

Lee, H. and Choi, C. (2022), The surprisingly straightforward scene text removal method with gated attention and region of interest generation: A comprehensive prominent model analysis, *in* 'European Conference on Computer Vision', Springer, pp. 457–472.

Li, C., Liu, W., Guo, R., Yin, X., Jiang, K., Du, Y., Du, Y., Zhu, L., Lai, B., Hu, X. et al. (n.d.), 'Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. arxiv 2022', *arXiv preprint arXiv:2206.03001* .

Zeng, W., Shu, Y., Li, Z., Yang, D. and Zhou, Y. (2024), 'Textctrl: Diffusion-based scene text editing with prior guidance control', *Advances in Neural Information Processing Systems* **37**, 138569–138594.