氏　　名	LEOW CHEE SIANG
博士の専攻分野の名称	博士（工学）
学 位 記 番 号	医工農博甲第１３６号
学位授与年月日	令和６年３月２２日
学位授与の要件	学位規則第４条第１項該当
専　　攻　　名	工学専攻　　システム統合工学コース
学 位 論 文 題 目	Studies on Text Detection and Character Image Generation for Advanced Text Recognition （テキスト認識高度化のためのテキスト検出 と文字画像生成に関する研究）
論 文 審 査 委 員	主査　教　授　西﨑　博光 教　授　鈴木　良弥 教　授　大渕　竜太郎 教　授　小澤　賢司 准教授　北村　敏也 教　授　伊藤　一帆

学位論文内容の要旨

In recent years, the digital landscape has seen a dramatic transformation, marking the onset of a new era in information accessibility and processing, largely driven by the swift advancements in deep learning technologies. This particularly impact in the realm of Optical Character Recognition (OCR), which has experienced a revolutionary change, echoing the evolution seen in the creation and distribution of multimedia content. Historically, OCR technology has faced significant challenges similar to those faced by early speech recognition systems, namely the accurate transformation of diverse textual content into machine-readable formats. These challenges were primarily due to the absence of powerful computational tools and sophisticated algorithms needed to manage the variability in text presentations. During this period, OCR systems were relatively basic and struggled with intricate layouts, various fonts, and inconsistent print quality, mirroring the early difficulties in speech recognition with unfamiliar words and different accents.

The introduction of deep learning marked a crucial turning point. The development and

widespread adoption of powerful Graphical Processing Units (GPUs), along with the expansion of data storage capabilities via Hard Disk Drives (HDDs) and Solid State Drives (SSDs), provided the necessary infrastructure for sophisticated computational tasks. This evolution in hardware, coupled with the exponential growth of big data, has propelled the advancement of deep learning technologies. Open-source deep learning frameworks like Tensorflow by Google and Pytorch by Meta (formerly Facebook), building on NVIDIA's Compute Unified Device Architecture (CUDA), have significantly enhanced OCR systems' capabilities. These advancements in OCR are reflective of the progress in speech processing, where deep learning has facilitated more efficient handling and interpretation of vast volumes of audio data.

Today's OCR technologies, powered by deep learning, demonstrate exceptional proficiency in accurately detecting and recognizing text from various sources. Modern systems are adept at handling multilingual text, deciphering handwritten notes, and processing documents with complex layouts, mirroring recent improvements in speech ecognition that enable nuanced understanding and interaction with humans. The field of OCR has benefited from methods such as single-line text detection in multi-line text blocks and innovative data augmentation techniques for character classification, improving the accuracy and reliability of these systems. The societal impact of these advancements in OCR technology is profound. In the business world, OCR systems have become essential for automating data entry, streamlining document management, and improving access to historical archives. In education, OCR facilitates the digitization of materials, making knowledge more accessible. In healthcare, the technology aids in managing patient records, enhancing the delivery of care.

Moreover, the integration of OCR with other technologies, like natural language processing and image recognition, opens up new avenues for advanced applications. For example, combining OCR with natural language processing and retrieval technologies can significantly improve information retrieval systems, making them more robust and user-friendly. The evolution of OCR, driven by deep learning and technological advancements, mirrors a broader trend in the digital age, where data processing and accessibility are constantly being redefined. As OCR technology continues to evolve, its integration with emerging technologies is expected to further revolutionize our interaction with and processing of the vast amounts of information available in our

increasingly digital world.

However, to build a high-performance OCR system with Deep Learning technology, a large number of data is required. OCR systems that currently exist in the world have high recognition rates for fonts that because of font training data can be generated easily. In contrast, handwritten text data must be written by hand by humans, which requires huge human and financial costs to generate large amounts of data. Today, there are far more documents containing not only machine printed characters but also handwritten characters than in the past, and there are all kinds of patterns of machine printed and handwritten characters, and OCR models based on Deep Learning are considered the most promising technology to handle them. In order to achieve highly accurate character recognition, it is also necessary to have a technology that can accurately detect characters. Due to the influence of digitization, text information printed on documents has become more complex, and it is difficult to accurately detect text because a large amount of text is printed on the commonly used A4 size paper, which is then further handwritten by humans. In particular, even if the characters are the same, they may be printed on multiple lines in a small area, making the boundary between characters ambiguous and making character detection more difficult.

The research objectives of this thesis include improving OCR accuracy using Deep Learning-generated training data, recognizing narrow multi-line characters more accurately, and developing methods for multi-line text recognition. The study focuses on the Y-Autoencoder (Y-AE) and CRAFT models, exploring their application in generating diverse character images and enhancing text detection accuracy. The research also aims to develop simpler approaches for recognizing characters in multi-line text environments, expanding the capabilities of deep learning models beyond traditional one-line recognition methods.

The thesis contributes to text recognition, text detection, and multiple-lines text recognition. It demonstrates that images produced by a deep learning model can enhance character image recognizer performance. The introduction of a novel post-processing method for existing deep learning models improves character recognition rates, especially for characters with narrow line spacing. The research also addresses limitations in conventional TrOCR systems, proposing a pre-processing technique for multi-line character recognition within TrOCR's fixed-size input

constraints.

論文審査結果の要旨

本博士論文では，ディープラーニング技術を用いたオプティカルキャラクター認識（OCR）の精度向上に焦点を当てており，大きく３つの主要なトピックに分かれている。

１つ目は，ディープラーニングに基づく生成モデルを使用したデータ拡張に基づく OCR 精度の向上である。この研究は，ディープラーニングによって生成された訓練データを活用して，OCR システムの文字認識精度を向上させる方法を提案している。これには、多様な文字イメージを生成し，これらを用いて OCR モデルの学習を行い，結果として認識精度を高めることができている。２つ目に，文書処理で必要な，狭い多行文字のより正確な文字認識に関する研究である。このトピックでは，狭い行間隔で印刷または手書きされた多行テキストを含む文書の認識精度を向上させることを目指している。特に，文字の境界が曖昧になりがちな小さな領域内の複数行にわたって印刷される文字の認識に焦点を当てており，これを正確に検出し認識する技術が研究された。最後は，多行テキスト認識に関する研究である。従来の OCR システムは一行テキストの認識に焦点を当てることが多いが，この研究トピックでは複雑な文書レイアウトにおける多行テキストの認識方法を研究することを目標としている。これには，多行テキスト環境での文字認識を簡素化する新しいアプリーチを提案している．公聴会および審査会においては，これらの内容について，発表が行われた。

まず，最初の生成モデルを用いたデータ拡張についてである。OCR の学習データ不足という課題に着目し，手書き文字画像生成モデルを用いて追加の学習データを自動生成することで、文字認識精度の向上を図る，というのが大きな目的である。本論文で提案した手書き文字画像生成モデルは Y-Autoencoder (Y-AE)と呼ばれるもので，これはエンコーダとデコーダから構成されたモデルである。Y-AE に画像のスタイル情報を反映できるように，"AdaIN"層を組み込むことで，入力画像の字体を模倣した多様な文字画像を生成できるように改良している点が研究のポイントとなっている。一方で，生成された文字画像の中にはノイズの多い不適切な画像，そもそも文字の形を成していないおかしな画像が大量に含まれる可能性があります。これをそのままデータ拡張につかっては，文字認識モデルの訓練に悪影響を及ぼします。そこで Mean Squared Error や事前学習済みの文字認識モデルに基づくフィルタリングを適用し，これらの不適切なノイズデータを除去する方法を提案しています。フィルタリング後の文字画像を用いて ResNet-152 ベースの文字認識モデルを学習することで，生成モデルを使わずに訓練した文字認識モデルと比較して，最大で

0.0494 ポイントの認識精度向上を実現した。これは文字誤認識率の改善率で，47.4%の精度改善に相当する。

次に，多行テキスト画像内の単一行テキストの検出精度を向上させるための研究について，発表が行われた。提案手法では，既存の文字検出モデルである"CRAFT"モデルを基にして，線によるテキスト分割を行うネットワークブランチをここに追加し，細かな多行セグメントの検出を可能にするモデルを新たに提案している。さらに，検出結果を後処理によってさらに高めて，より高度にテキストライン分割を実現することができている。実験では，異なる様々な文書形式から生成された文字画像データに対して，提案モデルの訓練と評価が行われた。評価データは，行間が狭い文字ライン画像を大量に用意して行った。実験では行間隔を変化させたテストデータを作成し評価した。結果として提案手法は従来手法と比べて，行間隔が狭い場合でもロバストな検出が可能で，OCR の文字認識精度向上にも有用であることが示された。

最後は，これまでの成果を活かして複数行の文字認識についての研究発表を行った。ここでは自己注意機構を用いた TrOCR というモデルを検討した。TrOCR は vision transformer と sequence transformer の 2 つのコンポーネントから構成される。Vision transformer が文字画像の特徴抽出を，sequence transformer がテキストのデコードを担っている。この構造により，複雑なテキスト認識に対して頑健な性能を発揮できる。実験では，TrOCR モデルを事前学習したデータセットでファインチューニングすることで，文字誤認識率(CER)が 6.54%まで改善されたことが確認できた。これは複雑な文字セットに対する認識精度向上を示している。さらに，論文の第 7 章で生成した文字画像データを用いた場合に，単一行文字では CER が 10%未満と良好な結果が得られた一方で，複数行文字ではモデルの学習が不安定になる場合があったことが判明したことについて報告した。これは生成データの取り扱いには注意が必要であることを示唆している。

本博士論文の主な成果をまとめると，1 点目は，Y-AE に AdaIN を統合することで，限られた学習データからでも多様な文字画像を生成できるようにした点が挙げられる。これにより OCR の学習データセットを拡張し, 文字認識精度の向上に寄与している。2 点目は，従来の CRAFT モデルに行分割ブランチを追加することで、複数行テキスト内の単一行テキストの検出精度を高めた点であり，これにより，行間隔の狭い複雑なテキストレイアウトでもロバストな文字検出が可能になった。最後は TrOCR モデルの入力画像に対して戦略的な前処理を施すことで，固定サイズの入力の制限にも関わらず，複数行の文字認識を実現した点であり，TrOCR の適用範囲を拡大させた。

博士論文公聴会および最終審査においては，審査員から， TrOCR の構造の改良や，構

造的に複雑な漢字に対しての頑健性，認識結果の曖昧さを中間的に評価できるかどうか，従来手法と比較した文字認識精度について質問があり，適切な回答がなされた。また，以上の研究内容について，査読付き国際会議論文，雑誌論文に採録されていることからも，本博士論文は，本学の博士学位を授与するに相応であると判断し，論文審査及び最終審査を合格とすることとした。