OPEN ACCESS

Manuscript ID:
TAM-08032024-7086

Volume: 8

Issue: 3

Month: January

Year: 2024

P-ISSN: 2454-3993

E-ISSN: 2582-2810

Received: 16.10.2023

Accepted: 21.12.2023

Published: 01.01.2024

Citation:
Nachiyar, Aranga. Kothai. "Data Testing in Digital Transmission: Tamil Literature." *Shanlax International Journal of Tamil Research*, vol. 8, no. 3, 2024, pp. 49–57.

DOI:
https://doi.org/10.34293/tamil.v8i3.7086

*Corresponding Author:
kothaiagp2008@gmail.com

Data Testing in Digital Transmission: Tamil Literature

Aranga. Kothai Nachiyar
Assistant Professor, Computer Science
Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India
https://orcid.org/0009-0005-6656-1120

Abstract

The developing area of Optical Character Recognition (OCR) is digital handwriting recognition. Manual writing is replaced by a digital writing pad. The font and shape of the letters vary while writing digitally. The writer's digital pen pressure and position on the digital pad cause covert text file problems during OCR recognition. When converting OCR to text, an error occurs because of the variations in letter shapes. In languages like Tamil, Chinese, Arabic, and Telugu, where the alphabet is made up of bends, curves, and rings, the aforementioned issue occurs. Tamil has more word mistakes in OCR-to-text conversion because the alphabet comprises curves and angles that must be correctly transcribed. The ResNet (Residual Neural Network) Two-Stage Bottleneck Architecture (RTSBA) is suggested in this paper. In order to recognize text written in Tamil on a digital writing pad, this article suggests using ResNet (Residual Neural Network) Two-Stage Bottleneck Architecture (RTSBA). The suggested RTSBA reduces the complexity of the Tamil alphabet recognition problem by using two distinct phases of neural networks. There are fewer inputs and variables in the early stages. Time and computational complexity are minimized in the last phase. A two-channel and two-stream transformer, long short-term memory, Inception-v3, recurrent neural networks, convolutional neural networks, and other conventional algorithms have been compared to the suggested algorithm. The digital writing pad-handwritten and HP lab datasets demonstrate the effectiveness of proposed methods like RTSBA, which yield accuracy rates of 98.7% and 97.1%, respectively.

Key Words: Optical Character Recognition, Handwritten Character Recognition, Tamil Language, Deep Learning, Techniques.

References

1. Raj, M. A. R., Abirami, S. and Shyni, S. M. *Tamil handwritten character recognition system using statistical algorithmic approaches*, Computer Speech & Language, vol. 78, Article ID 101448, 2023.
2. Jayanthi, V. and Thenmalar, S. *A review on recognizing offline Tamil manuscript character*, AIP Conference Proceedings, vol. 2591, no. 1, Article ID 020039, 2023.
3. Fateh, A. Rezvani, M. Tajary, A. and Fateh, M. *Persian printed text line detection based on font size*, Multimedia Tools and Applications, vol. 82, pp. 2393–2418, 2023.
4. Shanmugam, K. and Vanathi, B. *Newton algorithm based DELM for enhancing offline tamil handwritten character recognition*, International Journal of Pattern Recognition and Artificial Intelligence, vol. 36, no. 5, Article ID 2250020, 2022.
5. Fateh, A., Fateh, M. and Abolghasemi, V. *Multilingual handwritten numeral recognition using a robust deep network joint with transfer learning*, Information Sciences, vol. 581, pp. 479–494, 2021.
6. View at: Publisher Site | Google Scholar

மின்னணுப் பரிமாற்றத்தில் தரவுப் பரிசோதனை: தமிழ் மொழிஇலக்கியம்

அரங்க. கோதை நாச்சியார், **M.C.A., M.Phil.,**
உதவிப்பேராசிரியர், கணினித் துறை
அய்யநாடார் ஜானகி அம்மாள் கல்லூரி, சிவகாசி

ஆய்வுச்சுருக்கம்

ஆப்டிகல் கேரக்டர் ரெகக்னிஷனின் (OCR) *வளரும் பகுதி மின்னணு கையெழுத்து அங்கீகாரமாகும். கைமுறையாக எழுதுவது மின்னணு ரைட்டிங் பேடால் மாற்றப்படுகிறது. டிஜிட்டல் முறையில் எழுதும் போது எழுத்துக்களின் எழுத்துருவும் வடிவமும் மாறுபடும். எழுத்தாளரின் டிஜிட்டல் பேனா அழுத்தம் மற்றும் டிஜிட்டல் பேடில் உள்ள நிலை ஆகியவை* OCR *அங்கீகாரத்தின் போது இரகசிய உரை கோப்பு சிக்கல்களை ஏற்படுத்துகின்றன.* OCR *ஐ உரையாக மாற்றும்போது, எழுத்து வடிவங்களில் உள்ள மாறுபாடுகளால் பிழை ஏற்படுகிறது. தமிழ், சீனம், அரபு, தெலுங்கு போன்ற மொழிகளில், எழுத்துக்கள் வளைவுகள், வளைவுகள் மற்றும் வளையங்களால் ஆன, மேற்கூறிய சிக்கல் ஏற்படுகிறது.* OCR*-க்கு-உரையை மாற்றுவதில் தமிழில் அதிக வார்த்தைப் பிழைகள் உள்ளன. எழுத்துக்களில் வளைவுகள் மற்றும் கோணங்கள் உள்ளன. அவை சரியாக எழுதப்பட வேண்டும்.* ResNet (Residual Neural Network) *டூ-ஸ்டேஜ்* Bottleneck Architecture (RTSBA) *இந்தத் தாளில் பரிந்துரைக்கப்பட்டுள்ளது. டிஜிட்டல் ரைட்டிங் பேடில் தமிழில் எழுதப்பட்ட உரையை அடையாளம் காண, இந்தக் கட்டுரை* ResNet (Residual Neural Network) Two&stage Bottleneck Architecture (RTSBA) *ஐப் பயன்படுத்த பரிந்துரைக்கிறது. பரிந்துரைக்கப்பட்ட* RTSBA *ஆனது நரம்பியல் வலைப்பின்னல்களின் இரு வேறுபட்ட கட்டங்களைப் பயன்படுத்துவதன் மூலம் தமிழ் எழுத்துக்களை அடையாளம் காணும் சிக்கலின் சிக்கலைக் குறைக்கிறது. ஆரம்பக் கட்டங்களில் குறைவான உள்ளீடுகள் மற்றும் மாறிகள் உள்ளன. கடைசிக் கட்டத்தில் நேரம் மற்றும் கணக்கீட்டுச் சிக்கலானது குறைக்கப்படுகிறது. இரண்டு-சேனல் மற்றும் இரண்டு-ஸ்ட்ரீம் டிரான்ஸ்பார்மர், நீண்ட குறுகிய கால நினைவகம், இன்செப்ஷன்-வி3, மீண்டும் வரும் நரம்பியல் நெட்வொர்க்குகள், கன்வல்யூஷனல் நியூரல் நெட்வொர்க்குகள் மற்றும் பிற வழக்கமான அல்காரிதம்கள் பரிந்துரைக்கப்பட்ட அல்காரிதத்துடன் ஒப்பிடப்பட்டுள்ளன. டிஜிட்டல் ரைட்டிங் பேட்-கையால் எழுதப்பட்ட மற்றும் ஏக ஆய்வக தரவுத்தொகுப்புகள்* RTSBA *போன்ற முன்மொழியப்பட்ட முறைகளின் செயல்திறனை நிரூபிக்கின்றன. இது முறையே 98.7% மற்றும் 97.1% துல்லிய விகிதங்களை அளிக்கிறது.*

முக்கியச்சொற்கள்: *கணினி, டிஜிட்டல் பரிமாற்றம், தரவு, தமிழ்*

அறிமுகம்

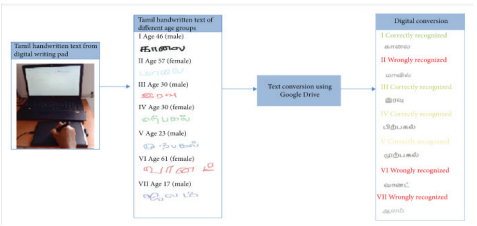
தமிழில் கையெழுத்தை அங்கீகரிப்பது தொடர் சிரமம். கன்வல்யூஷனல் நியூரல் நெட்வொர்க்குகளைப் பயன்படுத்தி, கையால் எழுதப்பட்ட தமிழ் உரையை (சிஎன்என்) அங்கீகரிப்பதற்காக ஆராய்ச்சியாளர்கள் வழிமுறைகளை உருவாக்கினர். அதிகரித்த அங்கீகாரத் துல்லியம் தேவை. தமிழ் என்பது கையெழுத்து அடிப்படையிலான மொழியாகும், இது பல்வேறு வழிகளில் எழுதப்படலாம், இது கையால் எழுதப்பட்ட தமிழ் எழுத்துக்களை துல்லியமாக அடையாளம் காணக் கடினமாக உள்ளது. கையெழுத்து அங்கீகாரத்தை மேம்படுத்த, எழுத்துப் பிரிவு மற்றும் எழுத்துரு இயல்பாக்கம் போன்ற பல்வேறு நுட்பங்களை ஆராய்ச்சியாளர்கள் பயன்படுத்துகின்றனர்.

தமிழ்க் கையெழுத்து சிக்கலான எழுத்துக்கள் மற்றும் வடிவங்களைக் கொண்டுள்ளது, இது கையால் எழுதப்பட்ட உரையை அடையாளம் காண்பதைக் கடினமாக்குகிறது. ஆழ்ந்த கற்றல் அடிப்படையிலான நுட்பங்கள், திரும்பத் திரும்ப வரும் நரம்பியல் நெட்வொர்க்குகள் (RNNPÒ) மற்றும் CNN கள், தமிழில் கையெழுத்து அடையாளத்திற்கான உயர்-துல்லியமான முடிவுகளை எவ்வாறு உருவாக்கக்கூடும் என்பதை அவை நிரூபிக்கின்றன. கூடுதலாக, கையால் எழுதப்பட்ட உரைப் படங்களில் இருந்து எழுத்துக்களை அகற்றும் அம்சம் பிரித்தெடுத்தல் மற்றும் படப்பிரிவைப் பயன்படுத்தி கையெழுத்து அங்கீகாரம் துல்லியம் அதிகரிக்கப்படுகிறது.

இருதரப்பு நீண்ட குறுகிய கால நினைவகம் (BiLSTM) குறியாக்கம் போன்ற கூடுதல் நுட்பங்கள், தமிழ்க் கையெழுத்தை மிகவும் துல்லியமாக அங்கீகரிப்பதற்காகப் பயன்படுத்தப்படுகின்றன.

ஆன்லைனில் கையால் எழுதப்பட்ட எழுத்துக்களில் உள்ள எழுத்துக்களை தானாக அடையாளம் கண்டுகொள்வதும், அதைத் தமிழ் எழுத்துக்கான ஆப்டிகல் கேரக்டர் ரெகக்னிஷனாக (OCR) மாற்றுவதும் கடினம். இணையத் தொழில்நுட்பத்தின் வளர்ச்சியுடன், டேப்லெட்களில் ஸ்டைலஸ் பேனா அல்லது விரல்களால் எழுதுவது மிகவும் பொதுவானதாகிவிட்டது. தமிழ்க் கையெழுத்தில் இருந்து உரைப் பிரிப்பு மற்றும் வகைப்படுத்தலுக்கு ஒரு புதிய அணுகுமுறை தேவைப்படுகிறது. அளவு மாறுபாடுகள், பரிமாண மாற்றங்கள், ஒழுங்கற்ற ஸ்டைலஸ் புள்ளிகள், கட்டமைப்புகளின் இடைநிறுத்தம், தேவையற்ற சுழல்கள், வடிவ ஏற்ற இறக்கங்கள் மற்றும் சிறப்பியல்பு வளைவுகள் ஆகியவை தமிழ்க் கையெழுத்து உரையின் முக்கிய சிக்கல்கள்.

ஆப்டிகல் டிஜிட்டல் கன்வெர்ஷன் டெக்னாலஜி மற்றும் பல்வேறு வயதினருக்கான கையால் எழுதப்பட்ட தமிழ் எழுதும் திண்டு படத்தில் சித்தரிக்கப்பட்டுள்ளது. டிஜிட்டல் பேனா மற்றும் பேடைப் பயன்படுத்திக் கையால் எழுதப்பட்ட காகிதத்தில் தவறான வார்த்தை அங்கீகாரத்தைக் காணலாம். ரெஸ்நெட் (எஞ்சிய நரம்பியல் நெட்வொர்க்) இரண்டு-நிலை பாட்டில்நெக் ஆர்கிடெக்சர் (RTSBA) மேற்கூறிய சிக்கலுக்குத் தீர்வாக வழங்கப்பட்டது. Residual Network (ResNet) எனப்படும் ஆழமான கற்றல் மாதிரி கணினி பார்வை பயன்பாடுகளில் பயன்படுத்தப்படுகிறது. ரெஸ்நெட்டின் இரண்டு-நிலை இடையூறு கட்டமைப்புச் சத்தத்தை வடிகட்டுதல் மற்றும் படத்தை மேம்படுத்துவதன் மூலம் பிணையச் செயல்திறனை மேம்படுத்தும் நோக்கம் கொண்டது.



பங்களிப்புகள்

1. டிஜிட்டல் பேட் மற்றும் பேனா உரை அங்கீகாரத்தில் கையால் எழுதப்பட்ட உரையை அங்கீகரிக்க, பிரிவை அடிப்படையாகக்கொண்ட RTSBA ஐப் பயன்படுத்துதல்.
2. பல்வேறு பேனா அழுத்தங்களில் டிஜிட்டல் பேட் மற்றும் பேனாவைப் பயன்படுத்திக் கையால் எழுதப்பட்ட உரையை அங்கீகரிக்க, பரிந்துரைக்கப்பட்ட RTSBA அல்காரிதத்தைப் பயன்படுத்த, இது எழுத்துக்களை எளிய வளைவுகளாகவும் மூடிய எளிய வளைவுகளாகவும் பிரிக்கிறது.
3. தமிழ் மொழியில் உள்ள தஞ்சாவூர், மதுரை, திருநெல்வேலி மற்றும் கோயம்புத்தூர் போன்ற தமிழ் மொழியில் உள்ள மக்கள்

தொகை அடிப்படையிலான டிஜிட்டல் எழுத்துக்களைக் கண்டறிதல், அங்கு தமிழ் எழுத்துக்களின் எளிமையான வளைவு எழுத்துக்களை மாற்றுகிறது.

4. தமிழில் மின்னணு எழுத்துகளை அடையாளம் காணும்பொருட்டு, எழுத்து அடிப்படையிலான எழுத்துக்கள் அங்கீகாரம் மற்றும் வகைப்படுத்தலுக்கு RTSBA ஐப் பயன்படுத்துதல்.
5. உட்பட பல்வேறு எழுதும் வயதுக் குழுக்களுக்கு ஏற்ப தமிழ் எழுத்துக்களை வகைப்படுத்தி அடையாளம் காணுதல் (i) **15-25,** (ii) **26-35,** (iii) **36-45,** (iv) **46-55,** மற்றும் (v) **56-65,** மற்றும் துல்லியம், நினைவுகூருதல் மற்றும் F1 மதிப்பெண்ணை ஒப்பிடுதல் பாரம்பரிய வழிமுறைகளுடன் முன்மொழியப்பட்ட முறை.

OCR தொழில்நுட்பத்தின் வரலாறு

இந்த ஆவணத்தை மாற்றும் தொழில் நுட்பத்தைக் கண்டுபிடித்த ரே குர்ஸ்வீல், **1974** இல் Kurzweil Computer Products, Inc. I நிறுவினார். ஏறக்குறைய எந்த எழுத்துருவிலும் தயாரிக்கப்பட்ட உரை இந்தப் புதிய தொழில்நுட்பத்தால் அங்கீகரிக்கப்படலாம். பார்வையற்றவர்களுக்கான இயந்திர கற்றல் கருவி அவரது கண்டுபிடிப்புக்கு மிகவும் பயனுள்ள பயன்பாடாக இருக்கும் என்று குர்ஸ்வீல் முடிவு செய்தார். உரையிலிருந்து பேச்சு மொழிபெயர்ப்பு மற்றும் சத்தமாக வாசிக்கும் திறன்கொண்ட வாசிப்புச் சாதனத்தை உருவாக்கினார். **1980** ஆம் ஆண்டில், அவர் தனது வணிகத்தை ஜெராக்ஸுக்கு விற்றார். ஏனெனில் பிந்தையவர் காகிதத்திலிருந்து கணினிக்கு உரை மாற்றத்தை விற்பனைச் செய்வதில் ஆர்வமாக இருந்தார்.

1990 களின் முற்பகுதியில் இந்தத் தொழில்நுட்பம் பிரபலமடைந்தது, ஏனெனில் இது பழைய செய்தித்தாள்களை டிஜிட்டல் மயமாக்கப் பயன்படுத்தப்பட்டது. அப்போதிருந்து, OCR இல் பல முன்னேற்றங்கள் ஏற்பட்டுள்ளன. இந்த நாட்களில், OCR ஆனது பயனர்களுக்குக் கிட்டத்தட்ட துல்லியமான மாற்றங்களை வழங்க முடியும். மேம்பட்ட OCR நுட்பங்கள் ஆவணச் செயலாக்க நடவடிக்கைகளின் தன்னியக்கத்தைச் செயல்படுத்துகின்றன. இந்த நிரல் கிடைக்கும் முன் ஆவணங்கள் கைமுறையாக மீண்டும் தட்டச்சுச் செய்யப்பட வேண்டும், இதற்கு அதிக நேரம், ஆற்றல் மற்றும் வளங்கள் தேவைப்பட்டன. இதன் விளைவாக, உள்ளடக்கச் சிக்கல்களின் வாய்ப்பு அதிகரித்தது. OCR இப்போது பரவலாகக் கிடைக்கிறது மற்றும் தனிப்பட்ட மற்றும் வணிகப் பயன்பாட்டிற்கு மிகவும் பயனுள்ளதாக இருக்கும்.

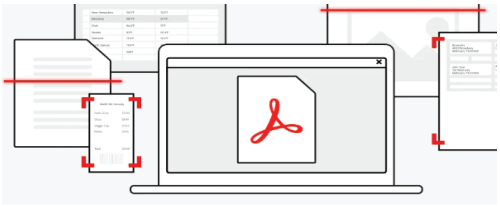
OCR தொழில்நுட்பத்தின் வகைகள்

தரவு விஞ்ஞானிகள் பல்வேறு வகையான OCR மென்பொருளை அவற்றின் பயன்பாடு மற்றும் பயன்பாட்டைப் பொறுத்து வேறுபடுத்துவார்கள். இதோ சில சான்றுகள்:

- வெவ்வேறு உரை மற்றும் எழுத்துரு பட வடிவங்கள் அடிப்படை ஆப்டிகல் எழுத்து அங்கீகார மென்பொருளால் டெம்ப்ளேட்களாகச் சேமிக்கப்படுகின்றன. பேட்டர்ன் டேஷ்-மேட்சிங் அல்காரிதம்களைப் பயன்படுத்துவதன் மூலம் மென்பொருளானது உரைப் படங்களுக்கு இடையே உள்ள மாறுபாடுகளை அடையாளம் காணும். அதன் உள் தரவுத்தளமானது தன்மைக்கு பாத்திரமாக ஆராயப்படும். உரையின் வார்த்தைக்கு வார்த்தைப் பிரதிபலிப்பு ஆப்டிகல் சொல் அங்கீகாரம் என்று அழைக்கப்படுகிறது. இந்த முறைக்கு வரம்புகள் உள்ளன, ஏனெனில் இது ஒவ்வொரு தட்டச்சு மற்றும் கையெழுத்து பாணியையும் பிடிக்க முடியாது, ஏனெனில் ஒவ்வொன்றிலும் எல்லையற்ற வேறுபாடுகள் உள்ளன.
- அறிவார்ந்த எழுத்து அங்கீகாரம் (ICR) மென்பொருள் நவீன OCR தொழில்நுட்பங்களின்

ஒரு பகுதியாகும். ஐசிஆர் உரையை மனிதர்கள் எப்படிப் படிக்கிறார்களோ அதே வழியில் படிக்கிறது. இயந்திர கற்றல் மென்பொருளைப் பயன்படுத்தி, இயந்திரங்களை மனிதர்களைப் போல செயல்பட பயிற்சி செய்யலாம். நியூரல் நெட்வொர்க் எனப்படும் இயந்திர கற்றல் அமைப்பு, உரையை ஆய்வு செய்து படங்களை மீண்டும்மீண்டும் செயலாக்குகிறது. இது கோடுகள், வளைவுகள், சுழல்கள் மற்றும் குறுக்கு வெட்டுகள் போன்ற பட அம்சங்களைத் தேடுகிறது மற்றும் இறுதி முடிவைப் பெற பல்வேறு நிலைத் தரவுகளின் முடிவுகளை ஒன்றாக இணைக்கிறது.

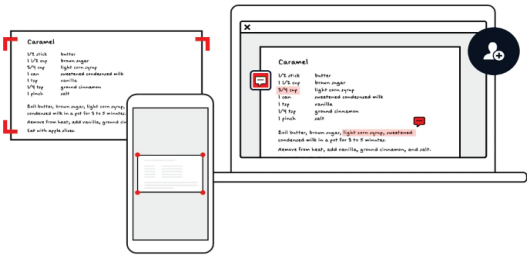
- புத்திசாலித்தனமான சொல் அங்கீகார தொழில்நுட்பங்கள் ICR போன்ற அதே விதிகளில் வேலை செய்கின்றன, ஆனால் அந்தத் தொழில்நுட்பங்கள் படங்களை எழுத்துகளாக மாற்றுவதற்குப் பதிலாக முழு வார்த்தைப் படங்களைப் படிக்கின்றன.
- ஆப்டிகல் மார்க் அறிதல் ஒரு ஆவணத்தில் வாட்டர்மார்க்ஸ், லோகோக்கள் மற்றும் பிற உரை அடையாளங்களைக் கண்டறியும்.



OCR மற்றும் இயந்திர கற்றல்

கடந்த சில தசாப்தங்களாக, OCR மற்றும் இயந்திர கற்றல் வேகமாக முன்னேறியுள்ளன, மேலும் வரும் ஆண்டுகளில், இரண்டு துறைகளும் சிறப்பாக இருக்கும். முந்தைய மென்பொருளின் தன்மையைப் பொருந்தும் அம்சங்களால் கட்டுப் படுத்தப்படுவதற்குப் பதிலாக, அடுத்த தலைமுறை OCR இன் வளர்ச்சியில் இயந்திர கற்றல் மற்றும் செயற்கை நுண்ணறிவு பயன்படுத்தப்படுகின்றன. ஓசிஆர் சாஃப்ட்வேர் தானே யோசித்து வளர்த்துக் கொண்டே போகிறது.

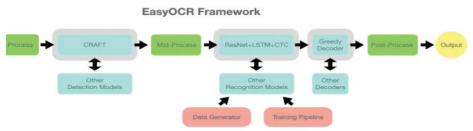
OCR தொழில்நுட்பம் ஸ்கேன் செய்யப்பட்ட உரையை விளக்கி, உரையைத் தொடர்ந்து அங்கீகரிப்பதோடு அதன் பொருளையும் புரிந்துகொள்ள முடியும். ஆழ்ந்த கற்றல் முன்னேற்றங்கள் மற்றும் OCR தொழில்நுட்பங்களை மாற்றுவதால், இயந்திரக் கற்றல் கடந்த காலத்தின் விஷயமாக மாறும். ஆழ்ந்த கற்றல் தொழில்நுட்பங்களில் பயன்படுத்தப்படும் நரம்பியல் நெட்வொர்க்குகள் மனித மூளையின் செயல்பாட்டைப் பிரதிபலிக்கின்றன, அல்காரிதம்கள் கடந்த கால வடிவங்களைச் சார்ந்து இல்லாமல் துல்லியமானவை என்பதைச் சரிபார்க்க முடியும். ஆழ்ந்த கற்றல் என்பது உரையைப் பார்க்கவும் அதன் அர்த்தத்தை அதன் சொந்தமாகக் குறைக்கவும் தொழில்நுட்பத்தின் திறனைக் குறிக்கிறது.



ஆழ்ந்த கற்றல் மற்றும் OCR மாதிரிகள், இயக்கவியல்

- ஆழமான கற்றல் முன்னேற்றங்களாக OCR சவாலுக்கான கூடுதல் தீர்வுகள் உருவாக்கப்படுகின்றன. அனலாக் உரையை டிஜிட்டல் வடிவத்திற்கு மாற்றுவது தற்போது பல்வேறு வழிகளில் செய்யப்படலாம். மிகவும் சுவாரசியமானவை பல ஆராயப்படும். முதலில் முக்கிய OCR செயலாக்கங்கள் மற்றும் OCR கடமைகளின் அளவைப் பார்ப்போம்.

OCR ஆழ்ந்த கற்றல் மாதிரியின் படிகள்



OCR அல்காரிதம் மூன்று அடிப்படை படிகளை உள்ளடக்கியது:

- உள்ளீட்டுப் படத்தை முன்கூட்டியே செயலாக்குகிறது. இந்த OCR படி எளிமைப்படுத்தல், அர்த்தமுள்ள விளிம்புகளைக் கண்டறிதல் மற்றும் உரை எழுத்துக்களின் வெளிப்புறத்தை வரையறுத்தல் ஆகியவை அடங்கும். எந்தவொரு பணிக்கும் இது ஒரு பொதுவான படியாகும், அதில் ஒரு படத்தை அடையாளம் காணும் கூறு உள்ளது. நீங்கள் ஆர்வமாக இருந்தால், படத்தைப் பற்றிய எங்கள் கட்டுரையில் இதே போன்ற அணுகுமுறையைப் பற்றி மேலும் விரிவாகப் பேசியுள்ளோம்.
- உரை கண்டறிதல். OCR திட்டப் பணியின் இந்தப் படிநிலைக்கு, படத்தில் காணப்படும் உரையின் துண்டுகளைச் சுற்றி ஒரு எல்லைப் பெட்டியை வரைய வேண்டும். SSD, நிகழ்நேர (YOLO) மற்றும் பிராந்திய அடிப்படையிலான கண்டறிதல்கள், நெகிழ் சாளர நுட்பம், மாஸ்க் R & CNN, EAST டிடெக்டர் போன்றவை இந்தப் படிநிலைக்கு பயன்படுத்தப்படும் பாரம்பரிய நுட்பங்களில் சில. (பட அங்கீகாரத்திற்கான ML மாதிரிகள் சிறப்பாகச் செயல்படவில்லை. உரையின் தனித்துவமான அம்சங்கள் காரணமாக OCR க்கு.)
- உரையின் அங்கீகாரம். இறுதி OCR படி, எல்லைப் பெட்டிகளில் வைக்கப்பட்டுள்ள உரையை அங்கீகரிப்பதாகும். இந்தப் பணிக்கு, ஒன்று அல்லது கன்வல்யூஷனல் மற்றும் ரிக்ரெண்ட் நரம்பியல் நெட்வொர்க்குகள் மற்றும் கவனம் செலுத்தும் வழிமுறைகளின் கலவை அடிக்கடி பயன்படுத்தப்படுகிறது. சில நேரங்களில் இந்தப் படிநிலை விளக்கப் படியையும் உள்ளடக்கியிருக்கலாம், இது கையெழுத்து அங்கீகாரம் மற்றும் IDC (புத்திசாலித்தனமான தரவுப் பிடிப்பு) போன்ற மிகவும் சிக்கலான OCR பணிகளுக்கான சிறப்பியல்பு ஆகும்.

அதற்குப் பதிலாக, OCR பணிகளுக்கு மிகவும் குறிப்பிட்ட உரை கண்டறிதல் மற்றும் உரை அங்கீகாரத்தின் கடைசி இரண்டு படிகளில் கவனம் செலுத்துவோம், மேலும் OCR எவ்வாறு ஆழமான கற்றலைப் பயன்படுத்துகிறது என்பதைப் பார்ப்போம்.

கன்வல்யூஷனல் ரீகரண்ட் நியூரல் நெட்வொர்க்குடன் (CRNN) ஆழ்ந்த கற்றல் OCR

படங்கள் OCR க்கு முன் செயலாக்கப்பட்ட பிறகு இந்த முறை இரண்டு படிகளைப் பின்பற்றுகிறது:

- அம்சங்களைப் பிரித்தெடுக்க கன்வல்யூஷனல் நியூரல் நெட்வொர்க் (CNN).
- டெக்ஸ்ட் கேரக்டர்களின் இருப்பிடம் மற்றும் மதிப்பைக் கணிக்க மறுநிகழ்வு நரம்பியல் நெட்வொர்க் (RNN).

CNN (Convolutional Neural Network)

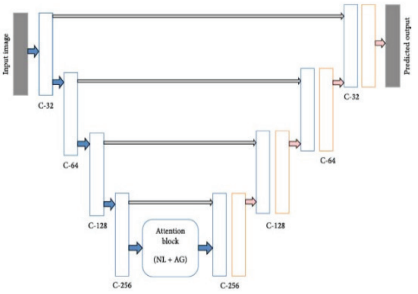
உரை கண்டறிதலுக்கான OCR அடிப்படையிலான ஆழ்ந்த கற்றலுக்கான சிறந்த முறைகளில் ஒன்று CNN ஆகும். கன்வல்யூஷன் லேயர்கள் அம்சங்களைப் பிரித்தெடுப்பதில் பயனுள்ளதாக இருப்பதால், அவை பட வகைப்பாடு பயன்பாடுகளுக்கு அடிக்கடி பயன்படுத்தப்படுகின்றன. ஒரு படத்தில் முக்கியமான விளிம்புகளையும், உயர் மட்டத்தில், சிக்கலான வடிவங்கள் மற்றும் பொருட்களையும் அடையாளம் காண்பதை அவை சாத்தியமாக்குகின்றன. எடுத்துக்காட்டாக, முழு-இணைக்கப்பட்ட அடுக்குகளுக்கு மாறாக, ஒரு படம் முழுவதும் பேட்டர்ன்-கண்டறிதல் வடிப்பான்களை மீண்டும் பயன்படுத்துவதன் மூலம், மாற்றியமைக்கும் அடுக்குகள், OCR அமைப்பின் சிக்கலைக் குறைக்கின்றன.

RNN (Recurrent Neural Network)

அடுத்தபடியாக RNNகளைப் பயன்படுத்தி பாத்திரங்களின் உறவுகளைத் தீர்மானிக்க வேண்டும். வெவ்வேறு நீளங்களைக் கொண்ட உள்ளீடுகளின் வரிசைகளைச் செயலாக்கும் போது, அத்தகைய பேச்சு அங்கீகாரம் அல்லது கட்டமைக்கப்படாத உரை (எ.கா., OCRக்கான கையெழுத்து அங்கீகாரம்), மீண்டும் மீண்டும் வரும் நெட்வொர்க்குகள் சிறந்து விளங்குகின்றன. மறைந்து வரும் சாய்வுச் சிக்கலைப் போக்க, நீண்ட குறுகிய கால நினைவகம் (LSTM) செல்கள் அடிக்கடி பயன்படுத்தப்படுகின்றன.

எளிய மற்றும் நேரான வளைவு உரை வகைப்பாட்டிற்கான RTSBA முறை

தமிழ் உரையை அங்கீகரிக்க இந்த ஆய்வில் பயன்படுத்தப்படும் பரிந்துரைக்கப்பட்ட RTSBA மாதிரி, கீழே உள்ள படத்தில் சித்தரிக்கப்பட்டுள்ளது. உரை காகிதத்தில் பேனாவால் எழுதப்பட்டு, ஆஃப்லைனில் குறிப்பு எடுப்பது மற்றும் இணைய ஆதாரங்கள் உட்பட பல ஆதாரங்களில் இருந்து சேகரிக்கப்படுகிறது. RTSBA பிரிவு இரண்டு தனித்தனி படிகள் மூலம் தமிழ்க் கையெழுத்து அங்கீகாரத்தின் சிரமத்தைக் குறைக்கிறது. நெட்வொர்க் முக்கியமான கூறுகளில் கவனம் செலுத்துவதால் மற்றும் முக்கியமற்ற பொருட்களைப்புறக்கணிப்பதால், இரண்டு-நிலை இடையூறு கட்டமைப்பானது உரையை துல்லியமாகப் பிரிக்கிறது. மேலும், துல்லியமான மற்றும் விரிவான பிரிவு விளைவுகளுக்கும் ஒற்றை-நிலை அமைப்புகளுக்கும் இடையே ஒரு ஒப்பீடு செய்யப்படுகிறது. RTSBA எழுத்துக்களின் வளைவுகளை மிகவும் சரியாகக் கணித்து, ஒவ்வொரு கட்டத்திலும் பல வகையான நரம்பியல் நெட்வொர்க்கு களைப் பயன்படுத்துவதன் மூலம் சிக்கலைக் குறைக்கிறது. RTSBA இரண்டு கட்டங்களாகப் பிரிக்கப்பட்டுள்ளது: குறியாக்கி மற்றும் குறிவிலக்கி, இடையில் கவனம் இடைவேளை. கவனத் தொகுதி NL மற்றும் AG தொகுதிகளுடன் ஒருங்கிணைக்கப்பட்டுள்ளது, இது CNN இல் உள்ள இடையூறுச் சிக்கல்களை சமாளிக்கிறது.



குறியாக்கிக் கட்டமானது கன்வல்யூஷன் லேயர்கள் மற்றும் மேக்ஸ் பூல் லேயர்களைக் கொண்டுள்ளது, அவை படங்களின் உள்ளடக்கத்தைப் பெறுகின்றன. கன்வல்யூஷன் அடுக்குகள் பட அம்சங்களைப் பிடிக்கின்றன, அதைத் தொடர்ந்து அம்ச அளவுருக்களை நீர்த்துப்போகச் செய்ய கீஉஃக் மற்றும் அதிகபட்சக் குளங்கள். அம்சம் இழப்பு மற்றும் பணிச்சுமைப் பிரச்சனைகளைச் சமாளிக்க, குறியாக்கிக் கட்டத்தில் சுருக்கம் மற்றும் தூண்டுதல் தொகுதிகள் அறிமுகப்படுத்தப்படுகின்றன. இரண்டு மடிப்பு அடுக்குகளுக்குப் பிறகு, படம் பிணையத்தில் நுழைகிறது; இரண்டு சேனல் பிரிப்பு பயன்படுத்தப்படுகிறது. ஒரு கீழ்-மாதிரி செயல்முறைப் படத்தை உள்ளீட்டு படத்தின் அதே அளவை உருவாக்குகிறது. மாதிரியானது விரிவான படப் பண்புகளைக் கற்றுக்கொள்கிறது, மேலும் அப்சாம்ப்லிங் செயல்முறையைப் பயன்படுத்திப் படத்தின் அளவு அதிகரிக்கிறது.

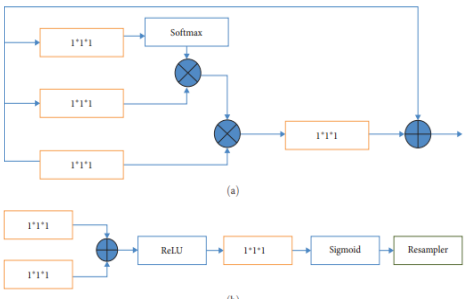
அம்ச வரைபடம் டிகோடர் கட்ட உள்ளீட்டுப் படத்தைப் போன்ற அளவிற்கு சரிசெய்யப்பட்டது. டிகோடர் கட்டம் நான்கு அப்சாம்லிங் தொகுதிகளைக் கொண்டுள்ளது. ஒவ்வொரு மாதிரித் தொகுதியும் இரண்டு மடிப்பு அடுக்குகளையும் ஒரு RELU அடுக்குகளையும் கொண்டுள்ளது.

குறியாக்கிக் கட்டத்தில் உள்ளீட்டுத் தகவலின் அளவு குறைகிறது மற்றும் குறிவிலக்கிக் கட்டத்தில் அதிகரிக்கிறது. உள்ளீடு சுருக்கப்பட்டு, இடையூறு ஏற்பட்டால் சில பண்புக்கூறுகள் அனுப்பப்படும். இடையூறுச் சிக்கல்களைச் சமாளிக்க, உள்ளீடு இழப்பைக் குறைக்க இரண்டு-நிலைத் தொகுதி உருவாக்கப்பட்டுள்ளது.

படம் a & b இல், NL தொகுதி அம்ச வரைபடம் குறிப்பிடப்படுகிறது; மற்றும் முறையே பெருக்கல் மற்றும் கூட்டலை சித்தரிக்கவும். ஒவ்வொரு வரிசையிலும், சமன்பாடு (1) இல் விவரிக்கப்பட்டுள்ளபடி, ஒவ்வொரு தொகுதியிலும் மென்மையான அதிகபட்ச செயல்பாடு செய்யப்படுகிறது.

$$Ci = Wzbi + ai \quad --------> 1$$

Wz ஆரம்ப எடை மதிப்புகளைக் குறிக்கிறது; ai என்பது எஞ்சிய தக வலுக்கானது; bi என்பது ஒத்த அளவு தக வலுக்கானது; மற்றும் ci என்பது தொகுதி மதிப்பிற்கானது. என்எல் மற்றும் ஏஜி பிளாக்குகளைப் பயன்படுத்தி இடையூறு பிரச்சினை தீர்க்கப்படுகிறது.



பரிசோதனை மற்றும் முடிவு

டிஜிட்டல் ரைட்டிங் பேட்-கையால் எழுதப்பட்ட (DWP&H) தரவுத்தொகுப்பு டேப்லெட் மாதிரி எண் Wacom CTL&672/ K0&CX ஐப் பயன்படுத்தி உருவாக்கப்பட்டது, இது ஆன்லைன் மற்றும் ஆஃப்லைன் பயன்பாட்டிற்கான கிராஃபிக் டேப்லெட் ஆகும். அழுத்தம் உணர்திறன் பேனாவுடன் டிஜிட்டல் பேடில் எழுதும்போது உரை தரவுத் தொகுப்புகள் சுற்றுப்புற விளக்குகளில் சேகரிக்கப்பட்டன. தி.டிஃப் பட வடிவத்தில், மொத்தம் 251 எழுத்தாளர் மாதிரிகள் பெறப்பட்டன. 10 முதல் 18 வயதுக்குட்பட்ட குழந்தைகள் முதல் 19 மற்றும் 59 வயதுக்குட்பட்ட பெரியவர்கள் மற்றும் 60 முதல் 75 வயதுக்குட்பட்டமூத்த ஆண்கள் மற்றும் பெண்கள் ஆகியோர் உரைப் படங்களின் அடிப்படையில் குழுவாக உள்ளனர். ஒவ்வொரு வகுப்பிலும், 92 X 133 எழுத்து அளவு கொண்ட 550 மாதிரிகள் இருந்தன, மேலும் ஒரு சிறிய சதவீதம் 10 வரை பங்களித்தது. மறுஅளவிடப்பட்ட படங்கள் பக்கவாட்டில் 50க்கு 50 பிக்சல்கள் நீளமாக இருக்கும். படத்தை இயல்பாக்க, இந்தப் படங்களில் உள்ள ஒவ்வொரு கிரேஸ்கேல் பிக்சல் மதிப்பும் 0, 1 வரம்பிலிருந்து -1, 1 வரம்பிற்கு மாற்றப்படுகிறது. ஆய்வுகளில், 0.0001 கற்றல் விகிதம் கொண்ட ஆடம் ஆப்டிமைசர் பயன்படுத்தப்பட்டது. நெட்வொர்க் 50 சகாப்தங்களில் 64 தொகுதி அளவுகளுடன் பயிற்சியளிக்கப்பட்டுள்ளது. முன்மொழியப்பட்ட மற்றும் RTSBA பின்வரும் அளவீடுகளின் அடிப்படையில் மதிப்பிடப்படுகிறது : தமிழில் கையால் எழுதப்பட்ட வார்த்தைகளுக்கான "துல்லியம்," "துல்லியம்,'' "ரீகால்," மற்றும் "F1 மதிப்பெண்".

தமிழில் இதேபோன்ற மற்ற எழுத்துக்கள் நீரூற்றுகள், கீழ் வளைவு மற்றும் வட்டம், நிற்கும் கோடு, கிடைக் கோடு, வலது, இடது, வட்டம், மேல், கீழ், புள்ளி, கேள்விக்குறி, சாய்ந்த கோடு மற்றும் நின்று, கிடை மற்றும் நிற்கும் கோடு. எழுத்துக்களில் மோதிரங்கள் இருப்பதால், முன்பு கூறப்பட்ட பக்கவாதம் இடையே வளைவுகளை வேறுபடுத்துவது சவாலானது.

தமிழ் கையால் எழுதப்பட்ட வார்த்தை அங்கீகாரம் CNN ஐப் பயன்படுத்தி செய்யப்படுகிறது. இது CNN மற்றும்

மாற்றியமைக்கப்பட்ட மல்டி-ஸ்கேல் செக்மென்டேஷன் நெட்வொர்க் (MMU & SNet) உடன் ஒப்பிடும்போது முன்மொழியப்பட்ட RTSBA இன் முடிவுகளைக் காட்டுகிறது. முதல் வார்த்தையில் உள்ள இரண்டாவது எழுத்து "O" வுக்குப் பதிலாக "K" என்று துல்லியமாகக் கணிக்கப்பட்டுள்ளது, மேலும் இந்த வார்த்தையின் பொருள் மாற்றப்பட்டது. இரண்டாவது வார்த்தையின் இரண்டாவது எழுத்து "ற" என்பதை விட "ந" என்று தவறாகப் புரிந்துகொள்ளப்படுகிறது. முதல் எழுத்து மூன்றாவது வார்த்தையை "சூ" என்பதைவிட "ஆ" என்று தவறாகக் கணித்துள்ளது. தமிழ் மொழியின் திறந்த-வளைவு மற்றும் மூடிய-வளைவு எழுத்துக்களின் செயல்திறன் அளவீடுகள் இங்கே சுருக்கப்பட்டுள்ளன.

Result Comparison CNN, MMU-SNet, and RTSBA

Table 2: Result comparison CNN, MMU-SNet, and RTSBA

Tamil handwritten word	Segmented	Correct word	CNN	MMU-SNet	RTSBA
		கழகு	✗	✓	✓
		பறவை	✗	✗	✓
		சூரி	✗	✗	✓

முடிவு மற்றும் எதிர்கால வேலை

டிஜிட்டல் ரைட்டிங் பேட் அடிப்படையிலான கையால் எழுதப்பட்ட எழுத்துக்களை அடையாளம் காணுதல், வகைப்படுத்துதல் மற்றும் சொல் அங்கீகாரம் ஆகியவற்றிற்கு, RTSBA பரிந்துரைக்கப்படுகிறது. பரிந்துரைக்கப்பட்ட RTSBA முறையானது, பேனா அழுத்தம், பேனா நிலை, எழுதும் பாணியில் உள்ள வேறுபாடுகள், வெவ்வேறு அளவுகளில் எழுத்துக்களுக்கு இடையே உள்ள இடைவெளிகள், எழுத்துக்களில் தேவையற்ற வளைவுகள், எளிய வளைவுகள், எளிய வளைவுகள், நேர்கோடுகள் ஆகியவற்றால் ஏற்படும் கையால் எழுதப்பட்ட எழுத்து மாற்றங்களைக் கண்டறியப் பயன்படுகிறது. திறந்த வளைவுகள் மற்றும் மூடிய வளைவுகள்.

பரிந்துரைக்கப்பட்ட RTSBA முறைகள் பாரம்பரிய அல்காரிதம்களுடன் ஒப்பிடும் போது, தோராயமாக 98.7% உரை கணிப்புத் துல்லியத்தைக் கொண்டுள்ளன. RTSBA அணுகுமுறையானது LSTM, Inception-v3, RNN, CNN, 2C2S மற்றும் MMU-SNet போன்ற பல நரம்பியல் நெட்வொர்க் வடிவமைப்புகளுடன் ஒப்பிடப்படுகிறது. மலையாளம் மற்றும் தெலுங்கு உள்ளிட்ட பிற இந்திய மொழிகளில் டிஜிட்டல் ரைட்டிங் பேடில் இருந்து உரை அங்கீகாரத்திற்கு இந்த அணுகுமுறை பயன்படுத்தப்படலாம்.

குறிப்புகள்

1. எம். ஏ. ஆர். ராஜ், எஸ். அபிராமி மற்றும் எஸ்.எம். ஷைனி, புள்ளியியல் அல்காரிதமிக் அணுகுமுறைகளைப் பயன்படுத்தி தமிழ் கையால் எழுதப்பட்ட எழுத்து அடையாள அமைப்பு, கணினி பேச்சு மற்றும் மொழி, தொகுதி. 78, கட்டுரை ஐடி 101448, 2023.
2. வி. ஜெயந்தி மற்றும் எஸ். தென்மலர், ஆஃப்லைன் தமிழ் கையெழுத்துப் பிரதியை அங்கீகரிப்பது பற்றிய ஆய்வு, AIP மாநாட்டுநடவடிக்கைகள், தொகுதி. 2591, எண். 1, கட்டுரை ஐடி 020039, 2023.
3. ஏ. ஃபதே, எம். ரெஸ்வானி, ஏ. தஜாரி மற்றும் எம். ஃபதே, பெர்சியன் அச்சிடப்பட்ட உரை வரி கண்டறிதல் எழுத்துரு அளவை அடிப்படையாகக் கொண்டது, மல்டிமீடியா கருவிகள் மற்றும் பயன்பாடுகள், தொகுதி. 82, பக். 2393-2418, 2023.
4. கே. சண்முகம் மற்றும் பி. வானதி, "நியூட்டன் அல்காரிதம் அடிப்படையிலான DELM ஆஃப்லைன் தமிழ் கையால் எழுதப்பட்ட எழுத்து அடையாளத்தை மேம்படுத்துவதற்கான,: சர்வதேச பேட்டர்ன் ரெகக்னிஷன் மற்றும் செயற்கை நுண்ணறிவு இதழ், தொகுதி. 36, எண். 5, கட்டுரை ஐடி 2250020, 2022.
5. ஏ.ஃபதே, எம்.ஃபதேமற்றும்வி.அபோல்கசெமி, "பரிமாற்ற கற்றலுடன் ஒரு வலுவான ஆழமான நெட்வொர்க் கூட்டுப் பயன்படுத்தி பன்மொழி கையால் எழுதப்பட்ட எண் அங்கீகாரம்," தகவல் அறிவியல், தொகுதி. 581, பக். 479-494, 2021.