Annual of Konstantin Preslavsky University of Shumen
Faculty of Mathematics and Informatics
vol. XXVI C, 2025, pp. 22 – 40, https://doi.org/10.46687/HIVE1644

ANALYSIS AND COMPARISON OF OPEN-SOURCE TOOLS FOR OPTICAL CHARACTER RECOGNITION IN BUSINESS DOCUMENT PROCESSING [*]

STOYAN K. KONSTANTINOV

ABSTRACT: *This paper compares leading open-source tools for optical character recognition (OCR) in business document processing. Five platforms—Tesseract, EasyOCR, PaddleOCR, docTR, and Donut—are evaluated for architecture, functionality, accuracy, and resource efficiency, with focus on Bulgarian-language support and intelligent document processing (IDP) capabilities such as layout, table, and key-value extraction. Findings show PaddleOCR offers the best balance of performance, usability, and cost, while hybrid OCR–language model approaches improve flexibility, and Donut fine-tuning achieves the highest accuracy for specific document types. Open-source OCR and IDP tools enable automation of document workflows while preserving data sovereignty and avoiding vendor lock-in.*

KEYWORDS: *Optical Character Recognition, Intelligent Document Processing, PaddleOCR, Open-Source Tools, Transformer*

2010 Math. Subject Classification: *68T05, 68U10, 68T45*

АНАЛИЗ И СРАВНЕНИЕ НА ИНСТРУМЕНТИ С ОТВОРЕН КОД ЗА ОПТИЧНО РАЗПОЗНАВАНЕ НА СИМВОЛИ ПРИ ОБРАБОТКА НА БИЗНЕС ДОКУМЕНТИ [†]

СТОЯН К. КОНСТАНТИНОВ

[*]This paper is (partially) supported by Scientific Research Grant РД-08-106/05.02.2025 of Shumen University

[†]Статията е частично финансирана по проект № РД-08-106/05.02.2025 на ШУ „Еп. Константин Преславски“

1 Въведение: От оптично разпознаване към интелигентна обработка на документи

Развитието на инструменти с отворен код за оптично разпознаване на символи (OPC), използвани при обработката на бизнес документи, се превръща в ключов елемент на дигиталната трансформация. Традиционните системи за OPC са създадени единствено да преобразуват изображение в машиночитаем текст, без да разбират структурата и семантичния контекст на документа [1]. За да се извлече реална бизнес стойност, са необходими решения от по-висок клас, известни като интелигентна обработка на документи (IDP – Intelligent Document Processing) [2], които комбинират три основни функции: анализ на оформлението, разпознаване на таблици и извличане на двойки ключ–стойност [3].

Исторически оптичното разпознаване има дълга еволюция, започваща още през 1960-те години, когато първите системи намират приложение в банковата и административната практика [1]. Ранните методи разчитат на шаблонно съпоставяне и морфологични операции върху изображенията, което ограничава точността и адаптивността им. Преходът към машинно и дълбоко обучение довежда до революция в областта: конволюционните и рекурентните невронни мрежи стават стандартни архитектури, а техният синтез в конволюционно-рекурентна невронна мрежа (CRNN – Convolutional Recurrent Neural Network) позволява постигане на значително по-висока точност и устойчивост на разпознаването [4].

През последното десетилетие фокусът се измества от чистото извличане на текст към по-дълбоко разбиране на документите. Интелигентната обработка на документи интегрира OPC с анализ на оформлението и семантичното съдържание [5], което дава възможност за автоматизирана обработка на фактури, договори и административни документи. Така се намалява нуждата от ръчна работа, ограничават се грешките и се ускоряват

процесите в различни сектори – банкиране, публично управление и здравеопазване [6].

Паралелно с технологичния напредък нараства интересът към отворените инструменти за OPC и IDP, които осигуряват прозрачност, адаптивност и независимост от конкретен доставчик. За българските организации това има особено значение, тъй като позволява поддръжка на български език и локално внедряване без зависимост от скъпи международни облачни услуги. В този контекст отворените технологии представляват не само икономически ефективна, но и стратегически устойчива алтернатива за изграждането на национална дигитална инфраструктура.

2 Базов стандарт: Tesseract за оптично разпознаване на символи

Tesseract, първоначално разработен от Hewlett-Packard през 80-те години и понастоящем поддържан от Google Research, е най-разпространеният инструмент с отворен код за оптично разпознаване на символи (OCR). Той представлява фундаментален еталон в развитието на OCR технологиите и служи като базова платформа за сравнение при оценка на съвременни системи. От версия 4 насам Tesseract използва архитектура, базирана на невронни мрежи с дълга краткотрайна памет (LSTM – Long Short-Term Memory), която моделира текстови последователности и позволява по-добра адаптация към вариации в шрифта, наклона и осветеността на изображението [7]. Това го отличава от класическите методи за разпознаване, разчитащи на шаблонно съвпадение или статистически анализ на символите.

Инструментът се разпространява под отворен лиценз Apache License 2.0, който допуска свободна употреба, модификация и интеграция в търговски решения [8]. Поддържа над сто езика, включително български, а езиковите пакети могат да бъдат дообучавани според конкретни нужди. За разлика от много съвременни решения, изискващи графични ускорители, Tesseract

функционира напълно върху централния процесор (CPU), което го прави достъпен за широк кръг организации и подходящ за внедряване в среди с ограничен хардуерен ресурс.

Архитектурата на Tesseract е изградена като последователен конвейер, който включва предварителна обработка на изображението, сегментация и разпознаване. В първия етап се прилагат техники за нормализация на контраста, корекция на наклона, бинаризация и премахване на шум с цел подобряване на четливостта. След това изображението се разделя на текстови региони чрез пространствен анализ, базиран на Connected Component Analysis (CCA), който идентифицира свързани групи пиксели и определя границите на символите. На следващ етап невронната мрежа LSTM разпознава символите в контекст, като запаметява последователността от входове и минимизира грешките, породени от сходни форми. Вграден класификатор за ориентация открива и автоматично коригира завъртян текст, което е особено важно при сканирани документи с отклонения от хоризонталата.

Основното ограничение на Tesseract се проявява при обработка на структурирани документи, като таблици, формуляри и фактури, при които липсата на модел за разбиране на оформлението води до възприемане на текста като линеен поток [9]. Системата не реконструира пространствените връзки между клетките и редовете, а извлича данните последователно, без да отчита тяхната позиция. Освен това точността ѝ зависи пряко от качеството на изображението, което налага внимателна предварителна обработка – изправяне, мащабиране, филтриране и контрастна корекция [11]. Средната точност при чист печатан текст е около 85%, но при документи със сложна структура тя намалява значително.

В експериментален сценарий с дигитализация на счетоводни книги от 70-те години, отпечатани върху пожълтяла хартия и с частично избледнял шрифт, Tesseract достига едва 45% точност без предварителна обработка. След прилагане на техники за

подобряване на контраста, запълване на липсващи пиксели и премахване на шум, точността се повишава до приблизително 78%. Това потвърждава ключовата зависимост между качеството на входните данни и крайния резултат от разпознаването.

Въпреки своите ограничения при сложни документи, Tesseract остава стабилен, надежден и напълно безплатен инструмент, който предлага висока точност при стандартни текстове и ниски изисквания за внедряване. Неговата отворена архитектура, поддръжката на множество езици и активната общност го превръщат в референтен модел и основа за развитие на съвременни системи за интелигентна обработка на документи (IDP), комбиниращи OCR с анализ на оформление и семантично разбиране.

3 Модерни решения: EasyOCR

EasyOCR, разработен от JaidedAI, представлява съвременен инструмент с отворен код за оптично разпознаване на символи, базиран на платформата PyTorch. Той се отличава с висока достъпност, лесна интеграция и способност за работа в мултиезична среда, като поддържа над осемдесет езика, включително български [12]. За разлика от класическия Tesseract, чиято архитектура изисква детайлна конфигурация и предварителна обработка, EasyOCR е проектиран с фокус върху бързото внедряване – системата функционира „готова за употреба“, без необходимост от обучение или ръчна настройка на параметрите. Издаден под свободния лиценз Apache 2.0, инструментът може да бъде използван както в академични, така и в комерсиални проекти, което го прави привлекателен за широк спектър от приложения в областта на дигиталната обработка на документи.

Архитектурата на EasyOCR е изградена върху двустепенен модел, съчетаващ откриване и разпознаване на текст. Първият компонент е CRAFT (Character-Region Awareness for Text Detection), който използва конволюционна невронна мрежа за

локализиране на областите, съдържащи текстови символи. Този детектор анализира изображението на ниво регион и извлича вероятностни карти, указващи къде е най-вероятно да се намират букви или думи. Вторият компонент е CRNN (Convolutional Recurrent Neural Network), който реализира самото разпознаване чрез комбинация от конволюционни и рекурентни слоеве. Конволюционните слоеве извличат пространствените характеристики на символите, а рекурентните моделират последователностите в текста. Благодарение на това EasyOCR постига добра устойчивост при обработка на текстове с различна ориентация, размер и шрифт, включително при снимки, сканирани документи и изображения с неравномерно осветление.

Въпреки своята гъвкавост и висока точност при чисти изображения, EasyOCR има ограничения при работа с документи, съдържащи сложни оформления. Системата не притежава вградени механизми за анализ на структурата на страницата и разглежда текста като линеен поток. Това я прави по-подходяща за извличане на изолирани текстови зони, но недостатъчно ефективна при документи с таблици, формуляри и взаимносвързани полета [13]. При подобни случаи резултатът е разпознаване на текста без възстановяване на логическата му организация, което ограничава употребата ѝ за интелигентна обработка на документи (IDP).

Друг важен аспект е производителността. EasyOCR е оптимизиран за работа с видеопроцесори (GPU), но може да функционира и на централния процесор (CPU) при по-ниска скорост. При масова обработка на документи GPU ускорението е почти задължително условие, тъй като осигурява многократно по-висока скорост на обработка. От друга страна, липсата на автоматични инструменти за предварителна обработка на изображението изисква използването на външни библиотеки като OpenCV или Pillow за подобряване на качеството на входните данни, особено при изображения с шум или нисък контраст.

Практическите експерименти показват, че при разпознаване на стандартни документи с висок контраст и добро качество EasyOCR достига точност над 90%, като успешно се справя с различни ориентации и смесени езици. При документи с неравномерна структура или изображения с артефакти точността спада, което потвърждава зависимостта на системата от качеството на входните данни и липсата на интегриран модул за анализ на оформление.

Въпреки тези ограничения, EasyOCR се утвърждава като надеждно и достъпно решение за бързо разпознаване на текст в мултимедийни и документообработващи системи. Неговата архитектурна простота, отворен лиценз и широк езиков обхват го правят предпочитан избор за изследователски и практико-приложни проекти, както и за системи, при които скоростта и лекотата на внедряване имат по-голямо значение от пълното структурно разбиране на документа.

4 Интегрирани платформи: PaddleOCR и docTR

Сред отворените инструменти за оптично разпознаване на символи и интелигентна обработка на документи особено място заемат PaddleOCR и docTR, които представляват ново поколение системи, надхвърлящи класическия подход към OCR чрез интеграция на структурен и семантичен анализ. Тези платформи комбинират висока точност, мащабируемост и възможност за внедряване в производствена среда, като се ориентират към пълната автоматизация на извличането на информация от бизнес документи [14][16].

PaddleOCR, разработен от Baidu AI Studio и базиран на платформата PaddlePaddle, е модулна система, която обединява няколко невронни компонента: детектор на текстови региони (DBNet), класификатор на ориентация и разпознавател с архитектура тип CRNN (Convolutional Recurrent Neural Network). Тази комбинация осигурява устойчиво и бързо разпознаване на текст при разнообразни условия на изображението. Особено

значимо е, че PaddleOCR поддържа над осемдесет езика, включително пълна поддръжка на български, и е издаден под лицензa Apache 2.0, което гарантира свободна употреба и разширяемост. За разлика от Tesseract и EasyOCR, PaddleOCR е оптимизиран както за графични ускорители (GPU), така и за централни процесори (CPU), като постига висока ефективност дори при големи обеми от данни.

Ключовото предимство на PaddleOCR е наличието на допълнителния модул PP-Structure, който трансформира системата от инструмент за OCR към пълноценна платформа за интелигентна обработка на документи (IDP). Този модул въвежда три фундаментални възможности: анализ на оформлението, извличане на таблици и разпознаване на двойки ключ–стойност. Чрез дълбоки модели за визуална сегментация PP-Structure разделя страницата на семантични региони — заглавия, основен текст, таблици, изображения или формули — като може да идентифицира повече от двадесет различни типа блокове. В табличните региони системата възстановява логическата структура на редовете и колоните, включително обединените клетки, което позволява точно възпроизвеждане на табличната информация и нейното преобразуване в машинночетим формат.

Особено интересна е възможността за семантично извличане на информация чрез двукомпонентен модел, съчетаващ Semantic Entity Recognition (SER) и Relation Extraction (RE). Тази комбинация позволява на системата да разпознава концептуални двойки ключ–стойност, дори когато те не са разположени последователно в документа, което е от решаващо значение при фактури, анекси и формуляри. Така PaddleOCR се приближава до функционалностите на комерсиалните IDP системи, като ABBYY FlexiCapture и Kofax, но при запазена отвореност и възможност за локално внедряване.

Производителността на PaddleOCR е оптимизирана чрез техники за квантизация и ускорено изчисление, позволяващи редуциране на размера на модела без съществена загуба на

точност. При работа на GPU (например NVIDIA A100) системата може да обработва до петстотин документа на час, докато на CPU производителността е по-ниска, но остава достатъчна за средни по мащаб внедрявания. Препоръчителната конфигурация за производствена среда включва видеопроцесор с 8–12 GB графична памет и използване на официалните Docker контейнери на PaddleOCR за стандартно и лесно поддържаемо внедряване [15].

Втората система – docTR, разработена от Mindee, представлява високопроизводителна библиотека, изградена върху платформата PyTorch, която използва съвременни архитектури с трансформатори за детекция и разпознаване на текст. Подходът при docTR е модулен – потребителят може да избира между различни модели за откриване и разпознаване според конкретното приложение. Това осигурява висока гъвкавост и възможност за интегриране в по-сложни системи за документообработка. Силната страна на docTR е скоростта на обработка и способността му да работи с документи в различни формати и ориентации, което го прави подходящ за реалновремеви приложения и обработка на потоци от изображения.

Основният недостатък на docTR е свързан с езиковата поддръжка. Предварително обучените му модели са ориентирани към латиница и френски език, като липсва „готова за употреба“ поддръжка на кирилица [16]. За да бъде използван ефективно при български документи, системата изисква процес на допълнително обучение (fine-tuning) с ръчно маркирани данни, което ограничава нейната достъпност за организации без капацитет за създаване на собствени корпуси. Въпреки това, docTR демонстрира отлична производителност при стандартни латински документи и е пример за следващото поколение OCR системи, използващи трансформаторни модели за визуално-текстов анализ.

Общият извод от анализа на двете платформи показва, че PaddleOCR се отличава с най-пълна функционалност и зрялост сред отворените решения, като предлага възможности за интеграция, анализ на оформление и извличане на структурирани

данни, докато docTR се фокусира върху скоростта и модулността, но е ограничен от езиковите си модели. И двете системи обаче представляват ключов етап в прехода от класическо OCR към интегрирани решения за интелигентна обработка на документи, които съчетават визуално разпознаване, семантичен анализ и висока производствена ефективност.

5 Следващо поколение: модели без оптично разпознаване

С развитието на дълбокото обучение и трансформаторните архитектури настъпва концептуален преход в областта на интелигентната обработка на документи – от класически системи за OCR към модели, които елиминират необходимостта от междинна фаза на текстово извличане. Тази нова група решения, известна като OCR-free подходи, не разпознава символи поотделно, а директно интерпретира визуалното съдържание на документа като структурирана информация. Най-представителен модел от този тип е Donut (Document Understanding Transformer), разработен от изследователската лаборатория Clova AI на NAVER Corporation, който реализира фундаментална промяна в начина, по който системите възприемат и обработват документи [17].

Donut е единен трансформаторен модел, който приема изображението на документа като вход и директно генерира структурирано текстово представяне на съдържанието под формата на JSON обект. За разлика от традиционните системи, състоящи се от два последователни етапа – откриване и разпознаване на текст, – Donut комбинира визуално възприятие и езиково моделиране в една архитектура. Тази цялостност елиминира грешките, натрупвани между отделните модули, и позволява на модела да разбира едновременно както текста, така и визуалната му организация. Архитектурата се състои от визуален енкодер, базиран на Swin Transformer, и текстов декодер, използващ архитектурата BART (Bidirectional and Auto-Regressive Transformer). Енкодерът анализира изображението като

йерархична мрежа от визуални токени, докато декодерът генерира текстови последователности, описващи съдържанието, логическата структура и контекста.

Основното предимство на Donut е способността му да комбинира визуално и семантично разбиране в един модел, което му позволява да извлича структурирани двойки ключ–стойност (KVP), заглавия, полета и таблици, без необходимост от предварително дефинирани шаблони. Така системата може да се справя с документи, които имат вариращи формати и непостоянно разположение на елементите, каквито са фактурите, застрахователните полици и административните формуляри. При правилно обучение Donut постига точност, сравнима или по-висока от тази на комбинации от OCR и езикови модели, тъй като елиминира натрупването на грешки между отделните компоненти.

Въпреки това, практическото внедряване на Donut изисква внимателен процес на допълнително обучение (fine-tuning). За да може моделът да интерпретира специфичен тип документи, е необходимо изграждане на обучаващ корпус, съдържащ между сто и петстотин ръчно анотирани примера, в които съдържанието е описано в структурирана JSON форма [18]. След това моделът се обучава да възпроизвежда тази структура при нови документи. Този процес гарантира висока точност при специализирани типове документи, но изисква значителни ресурси и експертиза в областта на машинното обучение. Размерът на модела и необходимата графична памет също са предизвикателство – оптималната работа на Donut изисква графични ускорители с 16–48 GB VRAM, което ограничава приложимостта му в среди с ограничен хардуер.

Развитието на мултимодалните големи езикови модели (VLM – Vision-Language Models), като Qwen-VL и PaLI-X, представлява следваща стъпка в тази посока. Тези модели могат едновременно да обработват изображения и текст, като изпълняват задачи за извличане на ключови полета, класификация или отговаряне на въпроси в т.нар. zero-shot режим – без специално обучение за конкретния тип документ [19]. Те притежават

способност да разбират визуалния контекст на страницата и да генерират логически свързани отговори, което отваря възможности за прилагане на естествен език при анализ на документи. Въпреки високия си потенциал, тези модели изискват огромни изчислителни ресурси и често разчитат на облачна инфраструктура, което ги прави по-малко подходящи за локално внедряване при ограничени бюджети или високи изисквания за защита на данните.

На този фон особено значим се оказва хибридният подход, който комбинира стабилността на традиционните OCR системи с гъвкавостта на езиковите модели. При него се използва трислоен процес: първо PaddleOCR извлича целия суров текст от изображението, след което този текст се подава към по-малък езиков модел, оптимизиран за скорост и обработка на естествен език. В последния етап моделът получава инструкция (prompt) да извлече конкретните двойки ключ–стойност или да структурира информацията според предварително зададен шаблон [20]. Тази комбинация позволява бърза адаптация към нови формати без необходимост от преобучение и минимизира изчислителните разходи, като запазва висока точност и гъвкавост.

Сравнителният анализ на Donut, мултимодалните модели и хибридните решения показва ясно разграничение между трите подхода. Donut постига най-висока точност при строго дефинирани и добре обучени задачи, мултимодалните модели предлагат универсалност без допълнително обучение, а хибридните решения осигуряват оптимален баланс между производителност, гъвкавост и разходи. В този контекст се очертава тенденция към конвергенция на двете парадигми – OCR и трансформаторните архитектури – в обща рамка, която интегрира визуално възприятие и езиково разбиране като основа за бъдещите системи за интелигентна обработка на документи.

6 Сравнителен анализ

Сравнителният анализ на разгледаните инструменти с отворен код за оптично и интелигентно разпознаване на документи показва отчетливо развитие от класически OCR системи към интегрирани и мултимодални архитектури, способни да обработват както визуалния, така и семантичния контекст на документа. Докато Tesseract и EasyOCR представляват зрели, стабилни решения за извличане на текст при ограничени ресурси, PaddleOCR, docTR и Donut демонстрират следващото поколение технологии, ориентирани към интелигентна обработка на документи (IDP). Разликите между тези системи се проявяват не само в точността на разпознаване, но и в начина, по който интерпретират оформлението, логическите зависимости и бизнес значението на съдържанието [14][16][17][20].

Tesseract остава най-достъпното решение, подходящо за архивни или административни документи с проста структура и високо качество на сканиране. EasyOCR предлага по-гъвкав и мултиезичен подход, но не поддържа анализ на оформление. PaddleOCR се отличава като най-зряло решение сред системите с отворен код – интегрира анализ на оформление, разпознаване на таблици и извличане на двойки ключ–стойност чрез модула PP-Structure, което го доближава до индустриалните решения за IDP. docTR допринася с висока скорост и модулност, но остава ограничен от липсата на готова поддръжка на кирилица. Donut, от своя страна, въвежда новата парадигма на OCR-free моделите, които елиминират междинното ниво на символно разпознаване и интерпретират документа директно като структурирано знание [17][19].

Количествените показатели от литературните сравнения и проведените тестове потвърждават, че PaddleOCR постига най-добър баланс между точност, производителност и изисквания към хардуера, като достига 94–97% при стандартни печатни документи и поддържа стабилна скорост на обработка дори при работа на

CPU. Donut демонстрира по-висока адаптивност при специализирани формати, но изисква процес на допълнително обучение (fine-tuning) и значителен графичен ресурс. Хибридните решения, комбиниращи PaddleOCR с малък езиков модел, предлагат практичен компромис – използват зрял OCR механизъм за извличане на текст, последван от езикова интерпретация на съдържанието чрез инструкции (prompt engineering), без необходимост от преобучение при промяна на формата [20].

Таблица № 1 Съпоставка на основните характеристики на инструментите

Характеристика	**Tesseract 5**	**EasyOCR**	**PaddleOCR**	**docTR**	**Donut**
Лиценз	Apache 2.0	Apache 2.0	Apache 2.0	Apache 2.0	MIT
Поддръжка на български (готов модел)	Да	Да	Да	Не (fine-tuning)	Да (с обучение)
Архитектура	LSTM	CRAFT + CRNN	DBNet + CRNN + PP-Structure	Transformer	Swin + BART
Анализ на оформление	Не	Ограничен	Отличен	Добър	Отличен
Извличане на таблици	Не	Не	Отличен	Среден	Отличен
Извличане на двойки ключ–стойност	Не	Не	Да (SER + RE)	Не	Да
Ресурсна интензивност	Ниска (CPU)	Средна (GPU препоръчителен)	Висока (GPU 8–12 GB)	Средна (GPU 6–8 GB)	Много висока (GPU 16–48 GB)
Средна точност (печатен текст)	80–85%	88–92%	94–97%	93–95%	96–98%

Характеристика	Tesseract 5	EasyOCR	PaddleOCR	docTR	Donut
Подходящ за структурирани документи	Не	Частично	Да	Частично	Да
Ниво на интеграция (IDP)	Ниско	Ниско	Високо	Средно	Много високо

Стратегически погледнато, PaddleOCR се очертава като най-прагматичният избор за реални внедрявания поради зрелостта си, баланса между точност и ефективност и пълната поддръжка на български език. Donut и свързаните с него трансформаторни модели представляват следващото поколение решения, насочени към визуално-езиково разбиране и по-дълбок контекстуален анализ. Хибридните системи комбинират предимствата на двата подхода, като съчетават надеждността на OCR с гъвкавостта на езиковите модели и представляват реалистично решение за организации с ограничени ресурси.

Отворените технологии за OCR и IDP вече демонстрират зрялост, сравнима с комерсиалните решения, като предоставят ефективни, мащабируеми и независими инструменти за автоматизирана обработка на документи. Тяхната стойност се изразява не само в икономическите параметри, но и в стратегическите предимства, свързани със суверенитета на данните, адаптивността и контрола върху внедряването. Тази еволюция очертава плавен преход към бъдещо поколение системи, при които визуалният и езиковият контекст се разглеждат като единно информационно пространство.

7 Заключение

Съвременните системи с отворен код за оптично и интелигентно разпознаване на документи демонстрират отчетлива еволюция – от класическо символно извличане на текст към интегрирани визуално-езикови архитектури, способни да

анализират структурата, оформлението и семантичните зависимости на документа. Тази трансформация представлява преход от техническо разпознаване към когнитивно разбиране, при което обработката на данни се превръща в част от интелигентни бизнес и административни процеси.

Сред разгледаните инструменти PaddleOCR се утвърждава като най-зрял и функционално завършен модел, който съчетава точност, стабилност и гъвкавост при внедряване. Donut и свързаните с него мултимодални трансформатори задават нова парадигма, базирана на визуално-езиково моделиране, което постепенно измества класическите OCR конвейери. Хибридните подходи, при които се комбинират OCR и езикови модели, предоставят балансирано решение между производителност, изчислителна ефективност и адаптивност към различни формати документи.

Използването на отворени инструменти има особено значение за българския контекст – те позволяват локално внедряване, поддръжка на български език и пълен контрол върху обработваните данни, без зависимост от чуждестранни облачни доставчици. Това осигурява както технологична независимост, така и съответствие с изискванията за защита на информацията.

Посоката на бъдещото развитие включва три основни тенденции: усъвършенстване на мултимодалните архитектури, оптимизация на ресурсите чрез компактни модели и разработване на езикови корпуси за по-малки езици, сред които българският. Тези направления потвърждават, че интелигентната обработка на документи с отворен код е не само технологична иновация, но и стратегически инструмент за дигитална трансформация, прозрачност и ефективно управление на информацията.

ЛИТЕРАТУРА

[1] Mori, S.; Nishida, H.; Yamada, H. Historical Review of OCR Research and Development. Proceedings of the IEEE, 1992, vol. 80, no. 7, pp. 1029-1058. Available at: https://ieeexplore.ieee.org

[2] Automated Table Extraction: OCR for Extracting Tabular Data. Klearstack, 2025. Available at: https://klearstack.com

[3] Bridging the Gap: OCR Techniques for Noisy and Distorted Texts. International Journal of Scientific Research in Computer Science Engineering and Information Technology, 2025, vol. 11, no. 1. Available at: https://ijsrcseit.com

[4] Smith, R. An Overview of the Tesseract OCR Engine. Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR), 2007, vol. 2, pp. 629-633. Available at: https://ieeexplore.ieee.org

[5] Subramani, N.; Akinade, O.; Gaur, P. A Survey of Deep Learning Approaches for OCR and Document Understanding. arXiv preprint arXiv:2011.13534, 2020. Available at: https://arxiv.org/abs/2011.13534

[6] Wang, X. F.; Jiang, X. D.; Zhang, Y.; et al. A Survey of Text Detection and Recognition Algorithms Based on Deep Learning. Science China Information Sciences, 2023, vol. 66, no. 3, pp. 131101. Available at: https://sciencedirect.com

[7] Graves, A.; Jäger, Y. Long Short-Term Memory Networks for Handwritten Character Recognition. IEEE Transactions on Neural Networks, 2009, vol. 20, no. 1, pp. 128-142. Available at: https://ieeexplore.ieee.org

[8] Kadaganchi, N.; et al. Optical Character Recognition using Tesseract Engine. International Journal of Engineering Research and Technology, 2021, vol. 10, no. 9. Available at: https://ijert.org

[9] Muthusundari, M. Optical Character Recognition System Using Artificial Intelligence and Machine Learning. Dialnet, 2024. Available at: https://dialnet.unirioja.es

[10] Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on

Computer Vision and Pattern Recognition, 2015, pp. 3431-3440. Available at: https://arxiv.org/abs/1411.4038

[11] Shi, B.; Bai, X.; Yao, C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, vol. 39, no. 11, pp. 2298-2304. Available at: https://ieeexplore.ieee.org

[12] Jain, V.; Seung, H. S. Natural Image Denoising with Convolutional Networks. Advances in Neural Information Processing Systems, 2008, vol. 21, pp. 769-776. Available at: https://arxiv.org/abs/0810.2408

[13] Hinton, G.; Osindero, S.; Teh, Y. W. A Fast Learning Algorithm for Deep Belief Nets. Neural Computation, 2006, vol. 18, no. 7, pp. 1527-1554. Available at: https://direct.mit.edu/neco

[14] Rosenberg, B.; Bin, H.; Thawonmas, R. Document Analysis and Recognition. Proceedings of the 18th International Conference on Document Analysis and Recognition (ICDAR), 2024, vols. 14804-14809 (LNCS). Available at: https://diva-portal.org

[15] Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1800-1807. Available at: https://arxiv.org/abs/1610.02357

[16] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961-2969. Available at: https://arxiv.org/abs/1703.06870

[17] Kim, G.; Hong, T.; Weissenborn, D.; Song, J. S. OCR-free Document Understanding Transformer. Proceedings of the European Conference on Computer Vision (ECCV), 2022, vol. 13688, pp. 498-517. Available at: https://arxiv.org/abs/2111.15664

[18] Devlin, J.; Chang, M. W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171-4186. Available at: https://arxiv.org/abs/1810.04805

[19] Radford, A.; Kim, J. W.; Hallacy, C.; et al. Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning, 2021, vol. 139 (PMLR), pp. 8748-8763. Available at: https://arxiv.org/abs/2103.14030

[20] Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention Is All You Need. Advances in Neural Information Processing Systems, 2017, vol. 30, pp. 5998-6008. Available at: https://arxiv.org/abs/1706.03762

Стоян Константинов
Direct Service Ltd.
e-mail: s.konstantinov@shu.bg