QO‘LYOZMALI HUJJAT TASVIRIDA MATN JOYLASHGAN SOHALARNI ANIQLASH USUL VA ALGORITMLARI

Mardiyev Azamat Shakar o‘g‘li
“Toshkent irrigatsiya va qishloq xo‘jaligini mexanizatsiyalash muhandislari instituti” Milliy tadqiqot universiteti huzuridagi Fundamental va amaliy tadqiqotlar institute
Email: mardievazamat7@gmail.com
Allayorov Jasur A’zamjon o‘g‘li
“Toshkent irrigatsiya va qishloq xo‘jaligini mexanizatsiyalash muhandislari instituti” Milliy tadqiqot universiteti huzuridagi Fundamental va amaliy tadqiqotlar institute
Email: jasur_allayorov@outlook.com
Alisherova Sarvinoz Alisherov qizi
“Toshkent irrigatsiya va qishloq xo‘jaligini mexanizatsiyalash muhandislari instituti” Milliy tadqiqot universiteti huzuridagi Fundamental va amaliy tadqiqotlar institute
Email: bilolbek02062019@gmail.com

Annotatsiya. Ushbu tadqiqotda hujjat tasvirlarida matn joylashgan sohalarni avtomatik aniqlashga oid usullar va algoritmlar keng qamrovda o‘rganildi. Ish davomida an’anaviy “pastdan yuqoriga” va “yuqoridan pastga” yondashuvlardan tortib, chuqur o‘rganishga asoslangan zamonaviy neyron tarmoqlar – to‘liq konvolyutsion tarmoq (FCN), U-Net va hujjatlarni tahlil qilish uchun e’tibor mexanizmiga asoslangan neyron tarmoq (DAN) arxitekturalarigacha bo‘lgan model va algoritmlar tahlil qilindi. Ayniqsa, tarixiy qo‘lyozmalardagi shovqin, og‘ish burchagi va bezakli elementlar mavjud sahifalar uchun moslashuvchan modellar ustunligi alohida yoritilgan. Tadqiqot natijalari chuqur o‘rganishga asoslangan va segmentatsiyasiz modellar an’anaviy algoritmlarga nisbatan aniqlik, tezlik va barqarorlik jihatidan sezilarli ustunlikka egaligini ko‘rsatdi. Natijada, bunday modellar hujjat tasvirlarida matn joylashgan sohalarni avtomatik aniqlash sohasida ilg‘or yondashuv sifatida e’tirof etilmoqda.

Kalit so‘zlar: qo‘lyozma matn, matn bloki, sahifa tuzilmasi, neyron tarmoq.

Аннотация: В данном исследовании всесторонне изучены методы и алгоритмы автоматического определения областей, содержащих текст, на изображениях документов. В ходе работы были проанализированы как традиционные подходы «снизу вверх» и «сверху вниз», так и современные нейронные сети, основанные на глубоком обучении – такие как полностью сверточная сеть (FCN), U-Net и нейронная сеть с механизмом внимания для анализа документов (DAN). Особое внимание уделено преимуществам адаптивных моделей, эффективных при обработке исторических рукописей,

содержащих шум, углы наклона и декоративные элементы. Результаты исследования показали, что модели, основанные на глубоком обучении и не требующие сегментации, значительно превосходят традиционные алгоритмы по точности, скорости и устойчивости. В итоге такие модели признаны передовым подходом в области автоматического определения текстовых областей на изображениях документов.

Ключевые слова: рукописный текст, текстовый блок, структура страницы, нейронная сеть.

Abstract: In this study, methods and algorithms for the automatic detection of text regions in document images were comprehensively examined. During the research, models and algorithms ranging from traditional “bottom-up” and “top-down” approaches to modern deep learning–based neural networks such as the Fully Convolutional Network (FCN), U-Net, and the Document Attention Network (DAN) were analyzed. Particular attention was paid to the advantages of adaptive models for pages containing noise, skew, and decorative elements in historical manuscripts. The results of the study showed that deep learning–based and segmentation-free models significantly outperform traditional algorithms in terms of accuracy, speed, and stability. Consequently, such models are recognized as advanced approaches in the field of automatic text region detection in document images.

Keywords: handwritten text, text blocks, page layout, neural network, page segmentation.

Kirish.

Axborot texnologiyalarining rivojlanishi bilan hujjatlarni raqamlashtirish va ularni avtomatik tahlil qilish muammosi dolzarb ahamiyat kasb etmoqda. Xususan, sahifa maketini avtomatik aniqlash – matn, jadval, tasvir va sarlavhalarni to‘g‘ri ajratish – hujjatni tushunish jarayonining muhim bosqichi hisoblanadi. Har bir hujjat – bu o‘ziga xos vizual tuzilma, unda matn, sarlavha, rasm va bezaklar uyg‘un holda joylashgan. Inson ko‘zi buni bir qarashda anglay oladi, ammo kompyuter uchun bu murakkab masaladir. Shu sababli olimlar yillar davomida sahifa maketini avtomatik tahlil qilish – ya’ni, sahifadagi matn maydonlarini, jadvallarni, sarlavhalarni va tasvirlarni aniqlash – uchun turli algoritmlar ishlab chiqishdi. Dastlabki usullar geometrik tahlilga asoslangan bo‘lsa, keyingi bosqichda chuqur o‘rganish texnologiyalari bu sohada yangi davrni boshlab berdi. Mazkur tadqiqot aynan shu o‘zgarishlarni yoritadi va hujjat tahlilining sun’iy intellekt algoritmlari yordamida qanday darajada takomillashganini ko‘rsatadi.

Tadqiqot maqsadi – qo‘lyozma hujjatlarining tuzilmasini avtomatik tahlil qiluvchi mavjud usul va yondashuvlarni tahlil qilib, qo‘lyozma matnni tanib olish

ISSN 3030-3907

uchun yaratilayotgan tizimga mos qo‘lyozmali hujjat tasvirida matn yozilgan hududlarni aniqlovchi samarali algoritmni tanlash.

Qo‘lyozma matn tasviridan matnlarni tanib olish jarayonida sahifada matn joylashgan o‘rinlarni avtomatik aniqlash muhim bosqich hisoblanadi. Qadimgi qo‘lyozmalarda matnlar sahifada bir nechta bloklarda yozilgan bo‘lishi hamda bloklardagi yozuvlar yo‘nalishlari ham turlicha bo‘lishi mumkin. Bundan tashqari sahifa chetki qismlarida naqshlar yoki bloklar o‘zaro naqshlar yordamida ajratilgan holatlar juda ko‘p uchraydi. Shuning uchun ham qo‘lyozma matnlarni tanib olish jarayonida sahifadagi matn joylashgan bloklarni aniqlash, sahifa tartibini avtomatik segmentatsiya qilish zarur. Qo‘lyozmali hujjat sahifasi tuzilmasini avtomatik tahlil qiluvchi usullarni to‘rtta asosiy guruh: “pastdan yuqoriga”, “yuqoridan pastga”, “neyron tarmoqqa asoslangan” va gibrid ga bo‘lish mumkin ekan.

Pastdan yuqoriga yondashuviga asoslangan usullar. [1] da sahifa maketini tahlil qilish uchun “docstrum” deb nomlangan hujjat spektri usuli taklif qilingan. Bu usul “pastdan yuqoriga” yondashuviga asoslangan va sahifa komponentalarini eng yaqin qo‘shnilar bo‘yicha klasterlashni amalga oshiruvchi sahifa maketini strukturaviy tahlil qilishga qaratilgan. U sahifa maketini tahlil qilishda yuqori darajada moslashuvchan. Bu usul yoradamida og’ish burchagi, qatorlar ichidagi masofalar, qatorlar orasidagi masofalar aniq o‘lchanadi hamda matn bloklari joylashuvi aniqlanadi. Ushbu usul og’ish burchagidan, turli xil shrift va oraliq masofalardan mustaqilligi hamda bitta sahifada bir nechta orientatsiyadagi bloklarni aniqlashi imkoniyatlari bilan boshqa usullardan afzallikka ega.

Yana bir an’anaviy “pastdan-yuqoriga” yondashuviga asoslangan hujjat tasviri maketini tahlil qiluvchi tezkor algoritm [2] da taklif qilingan. Ular asosan texnik jurnallar sahifalarini geometrik joylashuvi tahliliga e’tibor qaratgan. Ularning algoritmi hujjat tasviridan ulanma komponentlarni ajratib olishdan boshlanadi va ularni bloklar qo‘shniligi grafi orqali ifodalaydi. Matn orientatsiyasini baholash natijasida komponentlar umumlashtirilgan matn satrlariga birlashtiriladi. Mintaqalarni aniqlash geometrik chegaraviy qiymatlar va ierarxik modellashtirishga tayanadi, bu esa xususiyatlarni bir nechta darajada izchil ajratib olish imkonini beradi. Bu usul hujjat tasvirida matn, jadval, rasm va chizmalarni ajratishda juda samarali. Algoritm 95% dan yuqoriroq aniqlik ko‘rsatgan va bu algoritm katta hajmdagi hujjatlarni tahlil qilish uchun mos.

[3] da hujjat sahifalarini geometrik tahlil qilish uchun parametrsiz yondashuv taklif etildi. Avvalgi algoritmlar matn shrift hajmi, satrlar oralig‘i yoki hujjat tuzilishiga bog‘liq parametrlar talab qilgan bo‘lsa, ushbu usuldan foydalanishda bunday ehtiyoj yo‘q. Ushbu yondashuvda hujjat tasvirlarini

ISSN 3030-3907

maksimal darajada bir jinsli mintaqalarga – matnlar, jadvallar, tasvirlar va chiziqlar – ajratadi hamda oldindan belgilangan chegaraviy qiymatlarga tayanmaydi.

[4] da tarixiy qo‘lyozmalarning sahifa tuzilishini aniqlash uchun SIFT xususiyatlariga asoslangan yondashuv ishlab chiqilgan. Yondashuv binarizatsiyasiz ishlaydi va har bir sahifadagi joylashuv elementlarini – matn, sarlavha hamda bezakli bosh harflarni – lokal strukturalar asosida tahlil qiladi. DOG operatori yordamida aniqlangan qiziqish nuqtalar uchun SIFT tavsiflovchilari hisoblanib, SVM klassifikatori orqali asosiy matn va dekorativ elementlar ajratiladi. Matn satrlari esa DBSCAN klasterlash algoritmi yordamida segmentatsiya qilinadi. Yondashuv tarixiy qo‘lyozmalar uchun xos bo‘lgan shovqin, xiralik, notekis fon va o‘zgaruvchan yozuv uslublariga moslashuvchanligi bilan ajralib turadi. SIFT xususiyatlari tufayli metod masshtab, burilish va yorug‘lik o‘zgarishlariga nisbatan barqarorlikni ta’minlaydi.

Keyinchalik bir guruh olimlar qadimiy qo‘lyozmalarda matn joylashgan hududlar va matn qatorlarini segementatsiya qilish uchun o‘qitishdan holi bo‘lgan usulni taklif etdilar [5]. Bu usulning birinchi bosqichida Gabor filtri qo‘llanilib, matn teksturasidagi farqlar aniqlanadi hamda asosiy matn sohasi umumiy tarzda ajratiladi. Keyingi bosqichda esa grafik kesish algoritmi yordamida komponentlar “asosiy matn” va “yon yozuv” sifatida segmentatsiya qilinadi. Bu ikinchi bosqich energiyani minimallashtirish masalasi sifatida ifodalanadi. Energiyaning maqsad funksiyasi fazoviy yaxlitlik (spatial coherence) bilan bir qatorda komponent darajasidagi xususiyatlar, masalan, chiziq qalinligi, maydon va joylashuv kabi omillarni ham birlashtiradi. Natijada fazoviy jihatdan izchil segmentatsiya xaritalari hosil qilinadi. Bu usul mashinali o‘qitishga ehtiyoj sezmaydi, shu sababli o‘qitilmagan ma’lumotlarda ham yuqori samaradorlikni ko‘rsatadi.

An’anaviy ulanma komponentlarni agregatsiyalashga asoslangan yondashuvlardan farqli o‘laroq, [6] da tarixiy qo‘lyozma sahifalarining tuzilmasi – klassifikatsiya muammosi sifatida talqin qilinadi. Bu usul binarizatsiyasiz sahifani piksel darajasida tahlil qilib, har bir pikselning koordinata, rang va tekstura xususiyatlarini hisobga oladi. Rang komponentlari va ularning statistik ko‘rsatkichlari, shuningdek LBP va Gabor asosidagi tekstura tavsiflovchilari asosida xususiyatlar vektori shakllantiriladi. Eng muhim belgilar tezkor korrelyatsiyaga asoslangan filtrlash (Fast Correlation-Based Filter – FCBF) algoritmi yordamida tanlab olinadi, so‘ng piksellar SVM klassifikatori orqali “matn”, “fon”, “dekoratsiya” yoki “chegara” sinflariga ajratiladi. Ushbu yondashuv binarizatsiya bosqichini istisno qilgani sababli rangli va shovqinli sahifalarda ham aniqlikni saqlaydi. Rang, tekstura va koordinata belgilarining kombinatsiyasi sahifa strukturasini aniqroq ifodalab, segmentatsiya aniqligini oshiradi. Natijalar

ISSN 3030-3907

yuqori samaradorlikni ko‘rsatgan – turli qo‘lyozma to‘plamlarida 91-97% aniqlikka erishilgan.

[7] da esa hujjat tasvirlarini tahlil qilishda matn va matn bo‘lmagan sohalarni ajratish uchun bog‘langan komponentlar xususiyatlariga asoslangan o‘qitishga asoslangan yondashuv taklif etilgan. Yondashuvning asosiy g‘oyasi – har bir komponentning geometrik va statistik xususiyatlarini (o‘lcham, cho‘zilish, zarba kengligi, zichlik, Hue momentlari, joylashuv koordinatalari) tahlil qilish orqali ularni “matn” yoki “matn bo‘lmagan” sifatida tasniflashdir. Tasniflash Adaboost va qarorlar daraxti kombinatsiyasida amalga oshirilib, qo‘shni komponentlar kontekstini hisobga olish orqali aniqlik oshiriladi. Ushbu yondashuv piksel yoki blok asosidagi segmentatsiya usullariga nisbatan tezroq, barqarorroq va shovqinga kam sezuvchanligi bilan ajralib turadi. Post-processing bosqichi orqali qo‘shni komponentlar orasidagi o‘lcham va ehtimollik o‘xshashligi tahlil qilinib, xatoliklar kamaytiriladi.

Yuqoridan pastga yondashuviga asoslangan usullar. An’anaviy yondashuvlaridan biri bo‘lmish [8]da hujjat maketini tahlil qilish uchun tezkor “yuqoridan pastga” usuli taklif etilgan bo‘lib, u CLiDE tizimida amalga oshirilgan. Ushbu usulda ulanma komponentlar graf cho‘qqilari sifatida qaraladi hamda ular orasida yangi “maksimal masofa metrikasi” aniqlanadi. Kruskal algoritmi yordamida minimal xarajatli yoyuvchi daraxt quriladi va undan so‘zlar, qatorlar, bloklar hamda ustunlar hosil qilinadi. Hisoblash murakkabligini kamaytirish uchun evristik usullar va yo‘lni qisqartirish texnikasi qo‘llanilib, deyarli chiziqli samaradorlikka erishiladi. Mazkur xususiyatlari tufayli algoritm murakkab maketlarni, jumladan, matn va grafik elementlar aralash joylashgan ko‘p ustunli sahifalarni ham yuqori aniqlik va tezlik bilan samarali qayta ishlash imkonini beradi.

Neyron tarmoqqa asoslangan usullar. 2016-yilda tarixiy hujjat tasvirlari uchun superpiksel klassifikatsiyasi va o‘qitilmagan xususiyatlarni o‘rganishga asoslangan ilg’or yondashuv ishlab chiqildi [9]. Taklif etilgan yondashuv piksel darajasidagi belgilashni superpiksel darajasidagi yondashuv bilan almashtiradi. Bu esa hisoblash xarajatlarini keskin kamaytiradi va shu bilan birga segmentatsiya sifatini saqlab qoladi. Dastlab hujjat tasvirlari oddiy chiziqli iterativ klasterlash algoritmi yordamida ortiqcha segmentatsiyalanadi So‘ngra xususiyatlar to‘g‘ridan to‘g‘ri piksel intensivlik qiymatlaridan ko‘p qatlamli konvolyutsion avtokodlovchilar yordamida o‘rganiladi. O‘rganilgan xususiyatlar asosida tayanch vektorlarga asoslangan klassifikatori har bir superpikselni chegara, fon, matn bloki yoki bezak sinflaridan biriga ajratadi. Shundan so‘ng bog‘langan komponentlarga asoslangan silliqlashtirish bosqichi orqali klassifikatsiya natijalari takomillashtiriladi. Taklif etilgan yondashuv binarizatsiyaga, oldindan belgilangan

ISSN 3030-3907

qoidalarga yoki o‘qitilgan modelga tayanmaydi, bu esa uni yanada moslashuvchan va umumiy qiladi.

Ikki o‘lchovli konvolyutsion tarmoqlarga (2D-CNN) asoslangan hujjat tahlili yondashuvlaridan farqli ravishda, hujjat tuzilmasini avtomatik tahlil qilish uchun [10] da tezkor, bir o‘lchovli konvolyutsion neyron tarmoq (1D-CNN) asosidagi usul ishlab chiqildi. Ushbu usul hisoblash xarajatlarini minimallashtirgan holda matn, jadval va tasvir mintaqalarini aniqlash hamda tasniflashni maqsad qilgan. Usulning dastlabki bosqichida sahifa ketma-ketlik uzunligiga asoslangan algoritm va morfologik kengaytirish yordamida kontent bloklariga ajratiladi. Keyingi bosqichda esa har bir blok gorizontal va vertikal proyeksiya profillari asosida o‘qitilgan yangi 1D-CNN modeli yordamida tasniflanadi. Bu usul an’anaviy ikki o‘lchovli (2D) tasvir parchalaridan (100×100 piksel) bir o‘lchovli (2×100) ma’lumotlarga o‘tish orqali xotira sarfini va hisoblash murakkabligini 50 martagacha kamaytiradi, shu bilan birga 96,7% aniqlik bilan 2D-CNN asosidagi bazaviy modelga teng natijalarni beradi. Taklif etilgan usul 300ta ArXiv maqolalaridan tashkil topgan maxsus ma’lumotlar to‘plamida sinovdan o‘tkazilgan. Natijalar shuni ko‘rsatdiki, usul bosma ilmiy hujjat maketlari uchun yuqori darajada barqarorlikni namoyish etgan hamda matn, jadval va tasvir mintaqalarini kam xatolik bilan to‘g‘ri ajrata olgan.

An’anaviy qoidaga asoslangan yoki teksturaga yo‘naltirilgan metodlardan farqli o‘laroq, [11] chuqur o‘rganishga asoslangan model xususiyatlarni ajratish va tasniflashni to‘liq yakuniy tarzda amalga oshiradi hamda har bir pikselni quyidagi to‘rtta toifadan biriga ajratadi: fon, asosiy matn, izohlar va bezak. Model VGG-16 arxitekturasining o‘zgartirilgan shakliga asoslangan bo‘lib, unda chuqurroq qabul maydonlari hamda ko‘p darajali xususiyatlarni birlashtirish mexanizmi qo‘llaniladi. Bu esa global kontekst bilan bir qatorda nozik chiziq tafsilotlarini ham aniqlash imkonini beradi. Post-processing bosqichi esa segmentatsiya xaritalarini yanada takomillashtiradi – shovqinlarni bartaraf etish, mahalliy noto‘g‘ri tasniflarni tuzatish hamda matn ichida joylashgan bezak kabi ustma-ust mintaqalarni aniqlash amalga oshiriladi. Ushbu model DIVA-HisDB ma’lumotlar to‘plamida baholangan bo‘lib, piksel darajasidagi aniqlik 99% dan yuqori natijani ko‘rsatgan. Bu natija konvolyutsion avtokodlovchi asosidagi avvalgi modellarning 95,5% aniqligidan ancha yuqori bo‘lib, silliqlik va tuzilmaviy aniqlik jihatlarida sezilarli yaxshilanishga erishilgan. Mazkur tadqiqot natijasida FCN modeli tarixiy qo‘lyozma segmentatsiyasi uchun eng ilg‘or mezon sifatida tan olingan hamda chuqur o‘rganishga asoslangan usullarning qo‘lda ishlab chiqilgan va gibrid yondashuvlar ustidan ustunligini aniq namoyon etgan.

[12] da umumiy konvolyutsion neyron tarmoq asosidagi hujjat segmentatsiyasi uchun moslashuvchan tizim taqdim etganlar. Ushbu tizim sahifa

ISSN 3030-3907

ajratish), bazaviy chiziqni aniqlash, tuzilma tahlili hamda bezak yoki fotografik elementlarni ajratish kabi turli vazifalar uchun qo‘llaniladi. Tizim to‘liq konvolyutsion neyron tarmoq arxitekturasiga asoslanib, ResNet-50 ni kichrayuvchi tarmoq yo‘li sifatida, kengayuvchi (tiklovchi) dekoder yo‘lini esa U-Netga o‘xshash o‘tkazuvchi bog’lanishlar bilan birlashtiradi. Tarmoq chiqishida piksel darajasidagi ehtimollik xaritalari hosil bo‘ladi. Keyingi post-processing bosqichida esa chegaraviy qiymatlash, morfologik amallar va bog‘langan komponentlar tahlili asosida vazifaga bog‘liq natijalar aniqlashtiriladi.

[13] da tarixiy qo‘lyozma hujjatlari uchun sahifa segmentatsiyasi, matn satrlarini segmentatsiya qilish hamda bazaviy chiziqlarni aniqlashni bir vaqtning o‘zida bajarishga qodir bo‘lgan ko‘p vazifali to‘liq konvolyutsion tarmoq asosidagi modelni taklif etildi. Model VGG-16 tayanch arxitekturasini kengaytirib, kengroq qabul maydoniga ega bo‘lish uchun qo‘shimcha konvolyutsion qatlamlarni qo‘shadi. Shu bilan birga, to‘rtta tarmoqli chiqishga ega bo‘lgan ko‘p natijali tuzilma joriy etilgan bo‘lib, u mos ravishda sahifa mintaqalari (asosiy matn, izohlar, bezaklar), binarizatsiya, va “matn satrlarining markaziy chiziqlari” ni bashorat qiladi. Taklif etilgan model piksel darajasidagi klassifikatsiyaga asoslanib, DIVA-HisDB ma’lumotlar to‘plamida o‘qitilgan. U quyi darajadagi va yuqori darajadagi xususiyatlarni birlashtirish orqali yuqori aniqlikdagi segmentatsiyani ta’minlaydi. Evristik post-processing bosqichi esa mahalliy noto‘g‘ri tasniflarni tuzatish, ustma-ust mintaqalarni birlashtirish, hamda bazaviy chiziqlarni eng kichik kvadratlar usuli asosida moslashni amalga oshiradi. Eksperimental natijalarga ko‘ra, model sahifa segmentatsiyasi IoU ≈ 95,5%, matn satrlarini segmentatsiya qilish F1 ≈ 99,7%, bazaviy chiziq aniqlash F1 ≈ 99,6% aniqliklarni bergan.Bu tadqiqot natijasida ko‘p vazifali FCN asosidagi birinchi yakuniy yechimlardan biri ishlab chiqildi. Ushbu model tuzilma tahlilining uchta asosiy bosqichini sahifa ajratish, satr segmentatsiyasi va bazaviy chiziq aniqlashni – bitta yuqori samarali model doirasida birlashtirib, murakkab qo‘lyozma hujjatlarni tahlil qilishda yangi bosqichni boshlab berdi.

[14] tadqiqotda tarixiy arab qo‘lyozmalarining maket tahlili uchun Siamese neyron tarmog‘iga asoslangan chuqur o‘rganish algoritmi taklif etilgan. Algoritm binarizatsiyasiz ishlaydi va murakkab strukturalarga ega, egilgan, yon yozuvli yoki shovqinli sahifalarda ham yuqori aniqlikni ta’minlaydi. Algoritmning mohiyati – sahifani kichik o‘lchamdagi bo‘laklarga ajratib, ularning o‘zaro o‘xshashligini Siamese CNN orqali baholashdir. O‘xshashlik natijalari asosida bo‘laklar uch sinfga: asosiy matn, yon yozuv va fon sifatida klasterlanadi. Yakuniy bosqichda post-processing yordamida kichik xatoliklar tuzatiladi. Siamese CNN arxitekturasi 5 ta konvolyutsion qatlamdan iborat bo‘lib, AlexNet modeliga asoslangan. Tarmoq juft bo‘laklar orqali o‘qitilgan va ADAM optimizer yordamida moslashtirilgan.

ISSN 3030-3907

Sinov natijalarida model F-o‘lchov ko‘rsatkichi bo‘yicha yuqori natijaga erishgan: asosiy matn uchun 98.59% va yon yozuvlar uchun 96.89%. Bu qiymatlar MLP (95.02%) va FCN (95.00%) modellaridan ustun natijalarni ko‘rsatgan. Ushbu algoritm arab qo‘lyozmalari tahlilida “eng ilg‘or yondashuv” sifatida e’tirof etilgan bo‘lib, u yozuvning egilishi, qorayishi yoki fon dog‘lari mavjud bo‘lgan sahifalarda ham yuqori barqarorlikni namoyon etadi. Shuningdek, u binarizatsiya jarayonini istisno qilgan holda, to‘g‘ridan-to‘g‘ri rangli yoki kulrang tasvirlarda ishlov berish imkonini yaratadi.

[15] da tarixiy hujjatlarning vizual elementlarini – matn satrlari, bezaklar, rasmlar va chizmalarni – avtomatik ajratish uchun universal, o‘qitishsiz tizim ishlab chiqilgan. Taklif etilgan tizim sun’iy yaratilgan hujjatlar asosida o‘qitilgan to‘liq konvolyutsion neyron tarmoq modeliga asoslanadi. Tizimning asosiy g‘oyasi – haqiqiy annotatsiyasiz, faqat sintetik ma’lumotlar yordamida modelni o‘qitish va uni real tarixiy hujjatlarda bevosita qo‘llashdir. Tizim uch asosiy komponentdan iborat: SynDoc generatori tomonidan hosil qilingan sun’iy hujjatlar, ResNet-18 + U-Net arxitekturasiga asoslangan FCN tarmog‘i va oddiy post-processing bosqichi. Tizim har bir pikselni “matn chizig‘i” yoki “ilustratsiya” sifatida tasniflaydi hamda binarizatsiyasiz, to‘g‘ridan-to‘g‘ri rangli tasvirlarda ishlaydi. Sinov natijalariga ko‘ra, docExtractor turli hujjat turlarida (cBAD, IlluHisDoc, Mandragore) qayta o‘qitishsiz ham yuqori aniqlikka (F-o‘lchov ≈ 0.83–0.93, mIoU ≈ 76–87%) erishgan. Tizimning afzalligi – umumiy arxitektura asosida turli vazifalarda qayta foydalanish imkoniyati, real sahifalarda binarizatsiyasiz ishlov berish va GPU yordamida real vaqt tezligida natija olishdir. Natijalar Tesseract4 va Mask R-CNN kabi ilgari mavjud modellar natijalaridan sezilarli ustun.

Markewich va boshqalar [16] hujjat maketini tahlil qilishni semantik segmentatsiya muammosi sifatida qayta ko‘rib chiqdilar hamda segmentatsiya yondashuvi zamonaviy hujjatni jarayonida hamon samarali va kuchli usul ekanligini ta’kidladilar. Tadqiqotchilar Dense Article Dataset (DAD) nomli 450 ta ilmiy maqoladan iborat keng ko‘lamli ma’lumotlar to‘plamini taqdim etdilar. Ushbu to‘plam 43 ta yuqori aniqlikdagi maket sinflari bo‘yicha yorliqlangan bo‘lib, u hujjat tahlili uchun yangi mezon sifatida ishlab chiqilgan. DAD ma’lumotlar to‘plamida to‘rtta ilg‘or segmentatsiya arxitekturasi baholangan: DeepLabV3+, FastFCN, PSP-UNet va Gated-SCNN. Kichik va kam uchraydigan hujjat obyektlari bilan bog‘liq cheklovlarni bartaraf etish maqsadida mualliflar vaznlangan chegaraviy quti regressiya yo‘qotish funksiyasi ni taklif etdilar. Ushbu yondashuv segmentatsiya aniqligi va obyektni lokalizatsiya qilish aniqligini yaxshilashga imkon berdi. DAD va PubLayNet ma’lumotlar to‘plamlarida o‘qitilgan DeepLabV3+ tarmog‘i eng yuqori natijalarga erishdi – mos ravishda F1 = 96,26% (DAD) va F1 = 97,11% (PubLayNet) – hamda boshqa ilg‘or

ISSN 3030-3907

arxitekturalardan ustun natija ko‘rsatdi. Bundan tashqari, ularning qisman o‘qitilgan bosqichma-bosqich annotatsiya tizimi qo‘lda belgilash vaqtini 38% ga qisqartirdi, bu esa segmentatsiya asosidagi modellarni katta hajmdagi hujjat annotatsiyasi uchun amaliy jihatdan samarali ekanligini tasdiqladi. Mazkur tadqiqot semantik segmentatsiyani hujjat maketini tushunish sohasida markaziy va yuqori samarali yondashuv sifatida qayta tikladi.

An’anaviy obyekt aniqlash usullaridan farqli ravishda, faqatgina chegara qutilari chegaralariga tayanmasdan, hujjat tarkibiy qismlarini piksel darajasida aniqlash imkonini beruvchi murakkab hujjat maketlari uchun instansiya darajasidagi segmentatsiya yondashuvi taklif etildi [17]. Taklif etilgan yondashuv Mask R-CNN va Mask Scoring R-CNN arxitekturalariga asoslanib, obyektni aniqlash va hujjat komponentlarini segmentatsiya qilish jarayonlarini birgalikda amalga oshiradi. Ushbu yondashuv matn, jadval, tasvir, ro‘yxat va sarlavha kabi sahifa elementlarini aniqlaydi hamda ularni yuqori aniqlikda ajratib ko‘rsatadi. Model PubLayNet (zamonaviy hujjatlar) va HJDataset (tarixiy hujjatlar) kabi ikki yirik ma’lumotlar to‘plamida baholangan. Natijalarga ko‘ra, tizim mos ravishda o‘rtacha aniqlik (mean Average Precision, mAP) bo‘yicha 0.92 (PubLayNet) va 0.82 (HJDataset) natijalarga erishgan bo‘lib, bu ko‘rsatkichlar standart deteksiya asosidagi usullardan sezilarli darajada yuqori. Eksperimental natijalar shuni ko‘rsatdiki, instansiya darajasidagi segmentatsiya ustma-ust joylashgan yoki noto‘g‘ri shakldagi elementlarni aniqlashda yuqori aniqlikni ta’minlaydi va vizual jihatdan yaxlit maket namoyishini hosil qiladi.

[18] da segmentatsiyasiz qo‘lyozma hujjatlarni tanish uchun Hujjatlarni tahlil qilish uchun e’tibor mexanizmiga asoslangan neyron tarmoq (DAN) nomli yangi arxitektura ishlab chiqilgan. Ushbu arxitektura an’anaviy “matn qatorlari segmentatsiyasi → matnni tanib olish” bosqichlarini bekor qilib, butun sahifani bir butun vizual obyekt sifatida o‘qiy oladigan neyron tarmoqni taklif etadi. Arxitektura uch asosiy moduldan iborat: ResNet-34 asosidagi CNN enkoder vizual xususiyatlarni ajratadi, ikki o‘lchamli e’tibor mexanizmi sahifadagi muhim joylarga diqqatni jamlaydi, va Transformer asosidagi til modeli belgilar ketma-ketligini tahlil qilib, yakuniy matnni generatsiya qiladi. Model IAM, RIMES va READ kabi mashhur qo‘lyozma datasetlarda sinovdan o‘tkazilgan. Natijalar shuni ko‘rsatadiki, DAN yondashuvi CER (Character Error Rate) ≈ 3–5% va WER (Word Error Rate) ≈ 10–14% bilan mavjud CRNN va Transformer asosidagi segmentatsiyali tizimlardan ustun natijalarga erishgan. Segmentatsiya bosqichining yo‘qligi jarayonni tezlashtirgan va xatolik manbalarini kamaytirgan. Shuningdek, “eng muhim e’tiborli hududlar” lar orqali modelning qarorlari izohlanadigan ko‘rinishga ega bo‘lgan. Arxitekturaning muhim afzalliklari – notekis, egilgan yoki tutashgan yozuvlarda ham barqaror ishlash, yozuvchilarning

ISSN 3030-3907

turli uslublariga moslashuvchanlik va GPU yordamida parallel ishlov berish imkoniyatidir. Ushbu arxitektura qo‘lyozma hujjatlarni tanib olish sohasida segmentatsiyasiz hujjat o‘qish paradigmasini boshlab bergan eng ilg‘or ishlanmalardan biri sifatida e’tirof etiladi.

Gibrid usular. An’anaviy usullardan farqli o‘laroq, [19] da sahifa tuzilmasini avtomatik aniqlash uchun gibrid algoritm ishlab chiqildi. Muallif “pastdan yuqoriga” komponent tahlili va “yuqoridan pastga” strukturaviy yondashuvni birlashtirdi. Algoritmning markaziy g‘oyasi – matn elementlari orasidagi tabulyatsiya chegaralari orqali ustun chegaralarini aniqlashdir. Ushbu algoritm morfologik filtrlash, bog‘langan komponentlarni ajratish va tabulyatsiya chegaralari koordinatalarini aniqlash bosqichlarini o‘z ichiga oladi. Tadqiqot Tesseract OCR tizimiga integratsiya qilingan bo‘lib, ICDAR 2007 test to‘plamida 90% dan ortiq aniqlikka erishgan. Shuning uchun ham maqolada taklif etilgan algoritm murakkab sahifa maketlarini (gazeta, jurnal, reklama) tahlil qilishda yuqori aniqlik va tezlikni ta’minlovchi, real vaqtga yaqin ishlaydigan samarali yondashuv sifatida e’tirof etilgan.

An’anaviy segmentatsiya yoki binarizatsiyaga tayanuvchi usullardan farqli ravishda, [20] da gazeta va jurnallar kabi murakkab hamda rangli hujjatlarning sahifa tuzilmasini tahlil qilish uchun kontur klassifikatsiyasi va morfologik amallarni birlashtiruvchi gibrid algoritm taklif etildi. Taklif etilgan algoritm oldingi va orqa mintaqalarni birgalikda tahlil qilib, ustma-ust joylashgan ustunlar, matn va tasvir bloklarining segmentatsiyasini yaxshilaydi. Dastlab, orqa fon mintaqalari gorizontal va vertikal tuzilma elementlari yordamida morfologik ochish amali orqali qayta ishlanadi va shu orqali ustunlar hamda satr ajratgichlarining joylashuvi aniqlanadi. Keyinchalik oldingi mintaqalar inversiya qilinadi, va ularga konturga asoslangan klassifikatsiya tatbiq etiladi. Ushbu bosqichda belgining o‘rtacha balandligi va geometrik xususiyatlar asosida ramkalar, tasvirlar va matn bloklari aniqlanadi. Kichik konturlar ularning o‘lchamning balandlikka nisbatiga qarab to‘ldiriladi. Shundan so‘ng, kichik konturlar ketma-ket kengaytirish amali orqali birlashtirilib, yakuniy matn bloklari shakllantiriladi.

[21] da hujjat sahifalarini avtomatik segmentatsiya qilish uchun minimal gomogenlik algoritmi yoki eeng kam birxillikka asoslangan algoritm va adaptiv matematik morfologiyaga asoslangan gibrid algoritm taklif etilgan. Algoritm tasvirni kichik bloklarga bo‘lib, ularning piksel intensivligi farqlarini tahlil qilish orqali tarkiban bir xil hududlarni aniqlaydi. Keyinchalik, adaptiv morfologik operatsiyalar – ochish, yopish va kengaytirish – yordamida hudud chegaralari aniqlanadi hamda o‘xshash bloklar yagona mantiqiy sohalarga birlashtiriladi. Algoritm oddiy, lekin samarali tuzilishga ega bo‘lib, qat’iy emas, balki moslashuvchan morfologik elementlardan foydalanadi. Bu esa turli tipdagi

ISSN 3030-3907

hujjatlar, shu jumladan matnli, jadvalli va rasmli sahifalarda yuqori aniqlikka erishishni ta’minlaydi. Sinov natijalari 94% dan ortiq umumiy aniqlikni ko‘rsatgan: matn segmentatsiyasida ≈96%, tasvirni aniqlashda ≈92% natija qayd etilgan. Taklif etilgan algoritm hisoblash murakkabligi past bo‘lgani uchun real vaqt rejimida qo‘llanilishi mumkin. U murakkab tuzilishga ega arxiv hujjatlarida ham segmentlarning yaxlitligi va chegaralarning aniqligini saqlagan holda barqaror ishlaydi.

Xulosa. Bu tadqiqotda hujjat va qo‘lyozmalardagi matn joylashuvini aniqlash bo‘yicha turli usullar o‘rganilgan. Avvalgi yondashuvlar oddiy matematik va geometrik hisob-kitoblarga asoslangan bo‘lsa, keyingi yillarda sun’iy intellekt va neyron tarmoqlar yordamida yanada aniq natijalarga erishilgan. Ayniqsa, FCN va U-Net modellaridan foydalangan algoritmlar matn, jadval va rasm sohalarini aniq ajrata olgan. Eng yangi DAN modelida esa butun sahifa bitta tasvir sifatida o‘qiladi va segmentatsiya bosqichi umuman kerak bo‘lmaydi. Shu bois, bu tadqiqot hujjatlarni avtomatik tahlil qilish sohasida katta qadam bo‘lib, tarixiy hujjatlarni raqamlashtirish va tanib olishni yengillashtiradi. Kelgusida e’tibor segmentatsiyasiz, end-to-end tarzdagi tizimlar va transformer arxitekturalarining yanada optimallashtirilgan variantlarini ishlab chiqishga qaratilishi lozim. Yuqoridagi tadqiqotlarni tahlil qilgan holda tizim uchun [14]-tadqiqotdagi Siamese neyron tarmog’iga asoslangan chuqur o‘rganish algoritmi tanlab olindi. Chunki tizim eski o‘zbek yozuvidagi qo‘lyozmalarni tanib olish jarayonini amalga oshiradi. Ma’lumki, eski o‘zbek yozuvi arab grafikasiga asoslangan, hamda qo‘lyozmalarda shovqinlar, xalaqitlar ko‘p uchraydi. Bundan tashqari qo’lyozmalardagi sahifalar murakkab tuzilmalarga, egilgan, yon yozuvli matnlarga, bezaklarga ega.

FOYDALANILGAN ADABIYOTLAR

1. O‘Gorman L. The document spectrum for page layout analysis //IEEE Transactions on pattern analysis and machine intelligence. – 2002. – V. 15. – №. 11. – P. 1162-1173.

2. Jain A. K., Yu B. Document representation and its application to page decomposition //IEEE Transactions on pattern analysis and machine intelligence. – 1998. – V. 20. – №. 3. – P. 294-308.

3. Lee S. W., Ryu D. S. Parameter-free geometric document layout analysis //IEEE Transactions on pattern analysis and machine intelligence. – 2001. – V. 23. – №. 11. – P. 1240-1256.

4. Garz A., Sablatnig R., Diem M. Layout analysis for historical manuscripts using sift features //2011 International Conference on Document Analysis and Recognition. – IEEE, 2011. – P. 508-512.

ISSN 3030-3907

5. Asi A. et al. A coarse-to-fine approach for layout analysis of ancient manuscripts //2014 14th international conference on frontiers in handwriting recognition. – IEEE, 2014. – P. 140-145.
6. Chen K. et al. Page segmentation for historical handwritten document images using color and texture features //2014 14th International Conference on Frontiers in Handwriting Recognition. – IEEE, 2014. – P. 488-493.
7. Le V. P. et al. Text and non-text segmentation based on connected component features //2015 13th International conference on document analysis and recognition (ICDAR). – IEEE, 2015. – P. 1096-1100.
8. Simon A., Pret J. C., Johnson A. P. A fast algorithm for bottom-up document layout analysis //IEEE Transactions on Pattern Analysis and Machine Intelligence. – 2002. – V. 19. – №. 3. – P. 273-277.
9. Chen K. et al. Page segmentation for historical document images based on superpixel classification with unsupervised feature learning //2016 12th IAPR Workshop on Document Analysis Systems (DAS). – IEEE, 2016. – P. 299-304.
10. Oliveira D. A. B., Viana M. P. Fast CNN-based document layout analysis //2017 IEEE International Conference on Computer Vision Workshops (ICCVW). – IEEE, 2017. – P. 1173-1180.
11. Xu Y. et al. Page segmentation for historical handwritten documents using fully convolutional networks //2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). – IEEE, 2017. – V. 1. – P. 541-546.
12. Oliveira S. A., Seguin B., Kaplan F. dhSegment: A generic deep-learning approach for document segmentation //2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). – IEEE, 2018. – P. 7-12.
13. Xu Y. et al. Multi-task Layout Analysis for Historical Handwritten Documents Using Fully Convolutional Networks //IJCAI. – 2018. – P. 1057-1063.
14. Alaasam R., Kurar B., El-Sana J. Layout analysis on challenging historical arabic manuscripts using siamese network //2019 International Conference on Document Analysis and Recognition (ICDAR). – IEEE, 2019. – C. 738-742.
15. Monnier T., Aubry M. docExtractor: An offthe-shelf historical document element extraction //2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). – P. 91-96.
16. Markewich L. et al. Segmentation for document layout analysis: not dead yet //International Journal on Document Analysis and Recognition (IJDAR). – 2022. – V. 25. – №. 2. – P. 67-77.
17. Biswas S. et al. Beyond document object detection: instance-level segmentation of complex layouts //International Journal on Document Analysis and Recognition (IJDAR). – 2021. – V. 24. – №. 3. – P. 269-281.

ISSN 3030-3907

18. Coquenet D., Chatelain C., Paquet T. Dan: a segmentation-free document attention network for handwritten document recognition //IEEE transactions on pattern analysis and machine intelligence. – 2023. – V. 45. – №. 7. – P. 8227-8243.
19. Smith R. W. Hybrid page layout analysis via tab-stop detection //2009 10th International Conference on Document Analysis and Recognition. – IEEE, 2009. – P. 241-245.
20. Vasilopoulos N., Kavallieratou E. Complex layout analysis based on contour classification and morphological operations //Proceedings of the 9th Hellenic Conference on Artificial Intelligence. – 2016. – P. 1-4.
21. Tran T. A., Na I. S., Kim S. H. Page segmentation using minimum homogeneity algorithm and adaptive mathematical morphology //International Journal on Document Analysis and Recognition (IJDAR). – 2016. – V. 19. – №. 3. – P. 191-209.

ISSN 3030-3907