

VIDEO TASVIRLARDAN MUHIM MATNLARNI ANIQLASH VA SIMVOLLARNI AJRATISH

Tuychiyev Iqboljon Xasanboyevich

University of Management and Future Technologies magistranti

E-mail: Iqboljon.tuychiyev@mail.ru

ANNOTATSIYA

Maqolada video tasvirlardan muhim matnlarni ajratib olish uchun juda sodda yondashuv taklif etildi. Birinchidan, biz tasvirni qismlarga ajratdik va ularning har bir qismdagi zichligi tekshirildi. Ushbu usul juda qulay va keng qamrovli bo'lib, u yordamida turli xil tillarni, qo'lyozma va hattoki xiralashgan tasvirlarni aniqlash mumkin. Dastur bir vaqtning o'zida o'zbek kiril va o'zbek lotin harflarini tanish va o'qish xususiyatiga ega, ushbu usulning aniqligi 95% dan yuqoriligi keltirildi.

Kalit so'zlar: protsessor, tasvir, matn, qo'lyozma, optik skanerlash.

ABSTRACT

The article proposed a very simple approach for extracting important texts from video images. First, we divided the image into parts and checked their density in each part. This method is very convenient and comprehensive, allowing it to recognize different languages, handwriting, and even blurry images. The program has the ability to recognize and read Uzbek Cyrillic and Uzbek Latin letters at the same time, the accuracy of this method is higher than 95%.

Keywords: processor, image, text, handwriting, optical scanning.

АННОТАЦИЯ

В статье предложен очень простой подход для извлечения важных текстов из видеоизображений. Сначала мы разделили изображение на части и проверили их плотность в каждой части. Этот способ очень удобен и всеобъемлющий, что позволяет ему распознавать разные языки, почерк и даже размытые изображения. Программа имеет возможность одновременно распознавать и читать узбекскую кириллицу и узбекскую латиницу, точность этого метода выше 95%.

Ключевые слова: протсессор, изображение, текст, рукопис, оптическое сканирование.

So‘nggi yillarda hujjatlarni raqamlashtirish tendensiyasi paydo bo‘ldi. Dunyoning raqamlashuvi jarayonida qog‘ozga asoslangan hujjatlarni yanada qulayroq, qidirish va saqlash uchun raqamligiga aylantirish zarur. Optik belgilarni tanish usuli yordamida ushbu jarayonlarni yanada osonlashtirish mumkin. OCR¹-ni skanerlangan tasvirlarini mexanik yoki elektron konvertatsiya qilish deb atash mumkin. Yarim asrdan ziyod vaqt mobaynida ushbu sohada izlanishlar olib borilmoqda hamda hozirda ham davom etmoqda. Shuningdek zamonaviy OCRda belgilarni tanib olish darajasi hujjatda 99% sifatli va qo‘lda yozilgan hujjatlarda 90% dan yuqori. Buzilgan hujjatlar va kitoblar uchun OCR samaradorligi 80%. So‘nggi vaqtlarda tasvirlardan matn chiqarish uchun ko‘plab dasturlar qo‘llanilmoqda. Ushbu maqolada FAST² nuqtasi algoritmiga asoslangan juda sodda yondashuvdan foydalanilgan bo‘lib, bunda birinchidan, hujjat rasmini belgilangan o‘lchamdagi bir-biriga mos kelmaydigan kichikroq qismlarga ajratiladi va har bir qismdagi zichlikni FAST burchakni aniqlash texnikasi yordamida tekshiriladi. Zich qismlar matn bloklari deb belgilanadi va unchalik zich bo‘lmagan qismi rasm hududi deb belgilanadi. Keyingi qadamda qismlarni guruhlash uchun ularning ulanishini tekshiramiz, shundan keyin matn qismi rasmdan ajratilishi mumkin. Matn qismi rasmdan ajratilgandan keyin matnli hudud quriladi va u saqlanadi. OCR tizimlari turli xil tarkibiy qismlardan iborat bo‘lib, skanerlash, oldindan ishlov berish, segmentatsiyalash, xususiyatlarni ajratib olish, tasniflash va tanib olish, postga ishlov berish kabi ko‘plab bosqichlardan iborat. Oldindan ishlov berish bosqichida to‘siqni va tasvirdagi o‘zgarishlarni olib tashlashdan iborat. OCR oflayn yoki onlayn tarzda amalga oshirilishi mumkin. Onlayn tanib olish OCR protsessori belgisini ularga berilgan tarzda taniydi. Oflayn usulda protsessor ikkala hujjatni ham, qo‘lda yozilgan belgilarni ham taniy oladi, ammo oflayn rejimda tanib olish skanerlangan tasvirlarning sifatiga bog‘liq bo‘ladi.

Belgilarni tanib olish — bu kirish belgilarini oldindan belgilangan belgilar sinfi bo‘yicha tasniflash jarayoni hisoblanadi. Belgilarni tanib olish rasmlardagi matnni aniqlashda qo‘llaniladi. Matn skanerlangan tasvir, kamera orqali tushurilgan tasvir yoki qo‘lda yozilgan matn bo‘lishi mumkin.

A. Tasvirlardagi matn: amaliy dasturlarda skanerlangan tasvirlar mukammal emas, chunki rasmdagi ba’zi bir keraksiz tafsilotlar tufayli xalal beradigan qismlar bo‘lishi mumkin, bu esa tasvirdagi belgilarni aniqlashda buzilishga va effektivlikning kamayishiga olib kelishi mumkin.

¹ OCR (Optical Character Recognition) - bu Optik belgilarni aniqlash

² FAST (Features from Accelerated Segment Test)- bu tezlashtirilgan segment testidan olingan xususiyatlar



1-rasm. OCR uchun bosqichlar

Oldindan ishlov berish xalal beradigan qismlarni olib tashlashni o'z ichiga oladi (Gauss filtri, Gabor filtri kabi filtrlarni qo'llash orqali) va tasvirni rangli tasvir kabi to'g'ri konvertatsiya qilish, tasvirni keyingi ishlov berish uchun kulrang shkalaga yoki ikkilik tasvirga aylantirilishi mumkin. Xususiyatlarni chiqarib olish talab qilinadigan xususiyatni tan olishni o'z ichiga oladi. Tasniflash va tanib olish bosqichi bu jarayonni qazib olish bosqichidir. OCR jarayonini tugatgandan so'ng, dasturga qarab bir nechta qayta ishlash bosqichlarini amalga oshirish kerak, masalan, hujjatlarni meta-ma'lumotlar bilan belgilash (muallif, yil va h.k.) yoki OCR xatolari va imlo xatolarini tuzatish uchun hujjatlarni tasdiqlash kerak bo'ladi

OCR ustida hozirgi kunda ham izlanishlar olib borilmoqda va ushbu texnologiyada katta yutuqlarga erishish nazarda tutilgan. Bu usulning kelajakdagi ko'lami mobil qurilmalardagi OCR, qo'l yozuvlarini tanib olish, ingliz tilidan tashqari tillar (o'zbek kirill va o'zbek lotin) harflarini tanish, videodan rasmlarni olish va qayta ishlash, eski hujjatlarni qayta ishlash va qayta tiklash va boshqa ko'p narsalarni amalga oshirish ko'zda tutilgan.

B. Tasvirdagi matnlar: tasvirda matnlar, rasmlar va grafikalar kabi turli xil ma'lumotlar mavjud bo'lib, ushbu ma'lumotlarni qo'lyozma, eski, bosma hujjatlarni, jurnallarni skanerlash va rasmga tushirish orqali olinadi. Ko'p kontrastli, piksellar soni pastligi, ranglarning o'zgarib ketishi, murakkab fon va noma'lum matn rangi, o'lchamlari, joylashuvi, yo'nalishi va hakazo kabi skanerlangan hujjatlarni tanib olishda bir qancha muammolarga duch kelinadi. OCR tizimi yuqori sifatli bo'lsa ham, yuqorida muhokama qilingan muammolar tufayli tizim hali ham kerakli natijalarni bera olmaydi. Shuning uchun OCR jarayoni rasmning fon qismi toza bo'lsa hamda rasmda ortiqcha belgilar mavjud bo'lmasa yaxshi natija beradi.

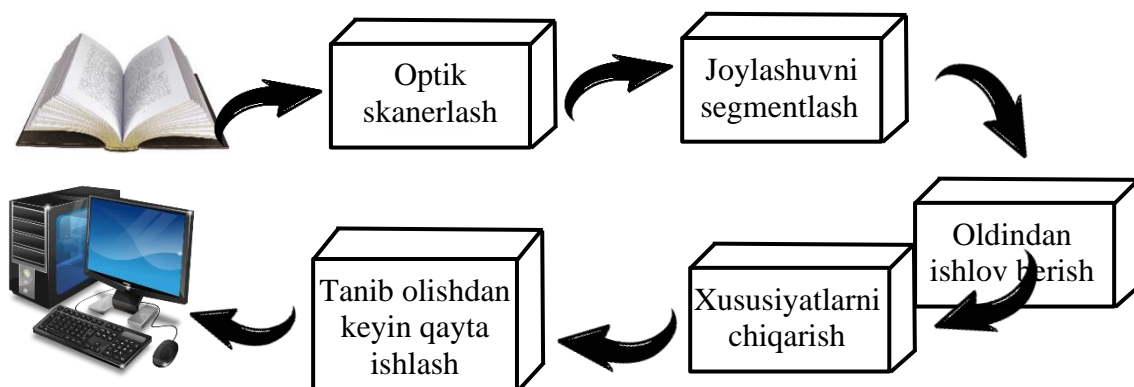
V. Rasmlardan matn hosil qilish: buning uchun ko'plab texnologiyalar qo'llanilgan. Ushbu maqolada FAST nuqtasi algoritmiga asoslangan sodda

yondashuvdan foydalanildi. Bunda birinchi rasmni belgilangan o'lchamdagi bir-biriga ustma-ust tushmaydigan kichik qismlarga ajratiladi. Keyin har bir blokda zichlikni FAST texnologiyasi asosida burchakni tekshiradi. Zich qismlar matn bloklari deb etiketlanadi va unchalik zich bo'lmagan rasmning qolgan qismi alohida blok sifatida olinadi. Keyin qismlarni guruhlash uchun qismlarning ulanishini tekshiramiz, shunda matnqismi rasmdan ajratilishi mumkin. Keyin biz matnli qismni quramiz va uni saqlaymiz.

Ushbu usul tez ishlaydigan va ko'p qirrali bo'lib, u yordamida turlixil tillarni, qo'l yozuvni va hattoki buzilgan hamda xiralashgan tasvirlarni aniqlashi mumkin. Bu sodda va qulay dastur bo'ib, ushbu usulning aniqligi 90% dan yuqori. Xulosa qilib aytganda, ushbu usul rasmlardan matnni aniqroq va murakkabroq aniqlashda yordam beradi.

OCR tizimining tarkibiy qismlari. Skanerlash bosqichida rasm hosil qilinadi va raqamlashtiriladi. Rasmning sifati ishlatilayotgan skanerga va rasmning asl nusxasiga bog'liq hisoblanadi. Amaliy dasturlarda skanerlangan tasvirlar mukammal emas, chunki rasmdagi ba'zi bir keraksiz tafsilotlar tufayli buzilish bo'lishi mumkin, bu esa tasvirdagi belgilarni aniqlashda xatoliklarga olib kelishi mumkin. Oldindan ishlov berish buzilishlarni olib tashlashni o'z ichiga oladi (Gauss filtri, Gabor filtri kabi filtrlarni qo'llash) va tasvirmirangli tasvir kabi to'g'ri konvertatsiya qilish tasvirni keyingi ishlov berish uchun kulrang shkalaga yoki ikkilik tasvirga aylantirilishi mumkin. Xususiyatlarni chiqarib olish talab qilinadigan xususiyatni tanolishni o'z ichiga oladi. Tasniflash va tanib olish bosqichi bu jarayonni qazib olish bosqichidir. OCR jarayonini tugatgandan so'ng, dasturga qarab bir nechta qayta ishlash bosqichlari zarur.

1. Optik skanerlash: bu jarayonda hujjatning raqamli tasviri olinadi. Hujjatlarni skanerlash uchun skanerdan yoki foto apparatdan foydalaniladi. Hujjatning sifati ishlatilayotgan skanerga va fotoapparatga bog'liqdir. Shunday qilib, tasvirni to'g'ri qayta ishlash uchun yuqori tezlik va yaxshi rang sifatiga ega skaner yoki foto apparat zarur.



2-rasm. OCR tizimining tarkibiy qismlari

Joylashtirish va segmentatsiya: ushbu jarayon tarkib mavjud bo'lgan joylarni topadi. Rasmning tarkibiy qismlarini aniqlaydigan jarayon – bu segmentatsiya deyiladi. Ma'lumotlar bosilgan hujjat qismini topish va ularni buzilgan nuqtalari va rasmlardan ajratish juda muhim jarayon hisoblanadi. Masalan, avtomatik ravishda pochta orqali saralash paytidamanzil tanib olish jarayonidan oldin joylashgan va marka yoki logotip singari konvertning boshqa tarkibiy qismlaridan ajratiladi.

Segmentatsiya – bu matnda bajariladigan belgilar yoki so'zlarni tasvirdan ajratish. Belgilarni aniqlashning aksariyat optik tizimlari so'zlarni alohida hujjatlashtirilgan ajratib olingan belgilarga ajratadi. Ushbu texnikani o'rnatish oson, ammo belgilar bir-biriga tegishi yoki belgilar bo'linib, bir nechta qismlardan iborat bo'lsa, muammolar paydobo'lishi mumkin. Segmentatsiyadagi asosiy muammolarni to'rt guruhga bo'lish mumkin:

1. Ta'sirchan va ajratilgan belgilarni chiqarish.
2. Matndan xalal beradigan nuqtalarni farqlash.
3. Matn uchun grafika yoki geometriyani xato bajarilishi. Bu matnni tanib olishda xatoliklarni keltirib chiqarishi mumkin.
4. Grafik yoki geometriya uchun matnni xato qilish. Bunday holda, matn tanib olish bosqichiga o'tkazilmaydi. Bu ko'pincha belgilar grafikaga ulangan bo'lsa sodir bo'ladi.

Taklif etilayotgan yondashuv: oddiy FAST algoritmidan foydalanib rasmlardan matnni ajratib olish yondashuvi taklif qilindi. Bu uchun birinchi navbatda tasvirni qismlarga ajratdik va ularning har bir qismdagi zichligi tekshirildi. Zich qismlar matn bloklari deb etiketlandi va unchalik zich bo'lmagan rasmning boshqa qismi xatolik qismlariga ajratildi. Matn qismini rasmdan ajratib olish uchun bloklarni guruhlab bog'lanishi aniqlandi. Ushbu usul oson va ko'p qirrali hisoblanib, bu yordamida turli xil tillarni ya'ni o'zbek lotin va o'zbek krill alifbosidagi matnlarni, qo'lyozmani hattoki buzilgan va xiralashgan tasvirlarni aniqlashimkonini bermoqda.

4-rasmda keltirilgan blok sxema taklif qilingan yondashuv asosidagi bosqichlarni ifodalaydi. Algoritmi quyidagi qadamlarda keltirilgan:

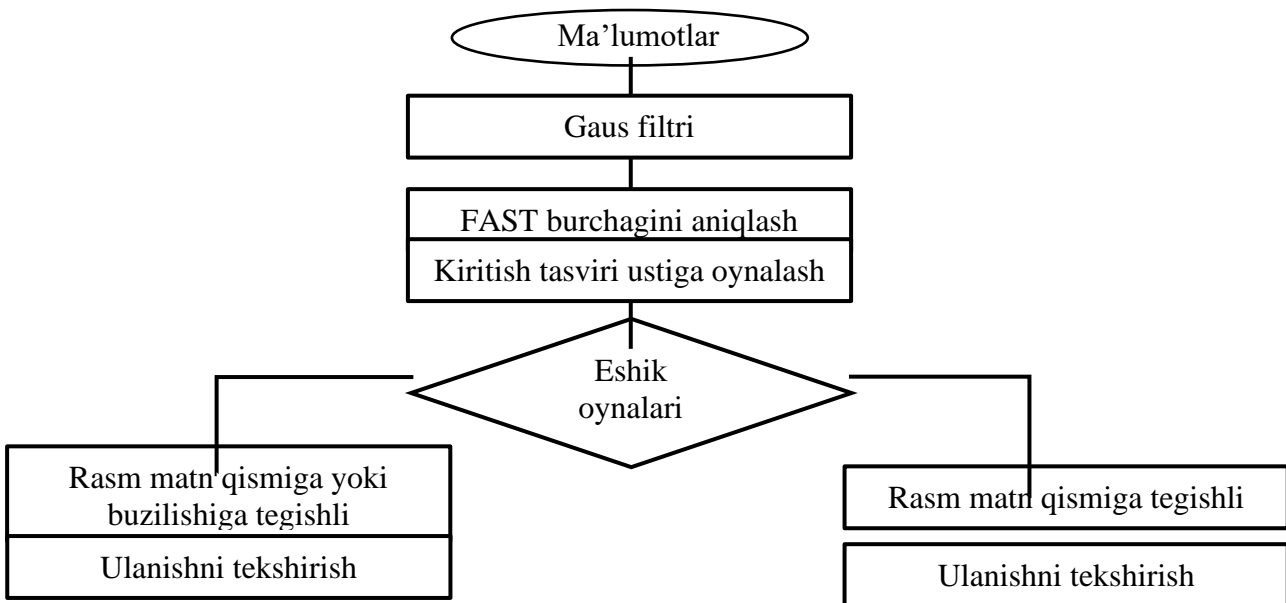
1-qadam: Rasm skanerlanadi va kulrang shkalaga aylantiriladi. Kulrang o'lchovli rasm ikkilik tasvirga aylantiriladi va tasvir raqamlashtiriladi ushbu usul (Binarizatsiya) deb nomlanadi. Maqolada Gauss filtridan foydalanilgan bo'ib, ushbu usul tasvirlarni xiralashtirish, xatoliklarni yo'qotish va rasmdagi keraksizdetallarni olib tashlash uchun ishlatiladi.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

2-qadam: Burchak nuqtalari FAST algoritmi bilan aniqlanadi

3-qadam: Tasvirni bir-biriga mos kelmaydigan qismlarga ajratib, burchak nuqtalari soni hisoblanadi.

4-qadam: Qismda maksimal burchak nuqtalariga (N_{max}) ega bo'lgan blok topiladi, tanlangan blok yordamida chegara aniqlang, $T = 0,2 * N_{max}$ sifatida ishlatiladigan chegara aniqlanadi. (Maksimal qiymatdan 20%).



3-rasm. Algoritm uchun oqim jadvali.

5-qadam: Burchaklar soni ko'proq bo'lgan bloklarni matnli qismlarga, chegarasi kamroq bo'lgan bloklar esa rasm yoki fon qismlarga taqsimlanadi.

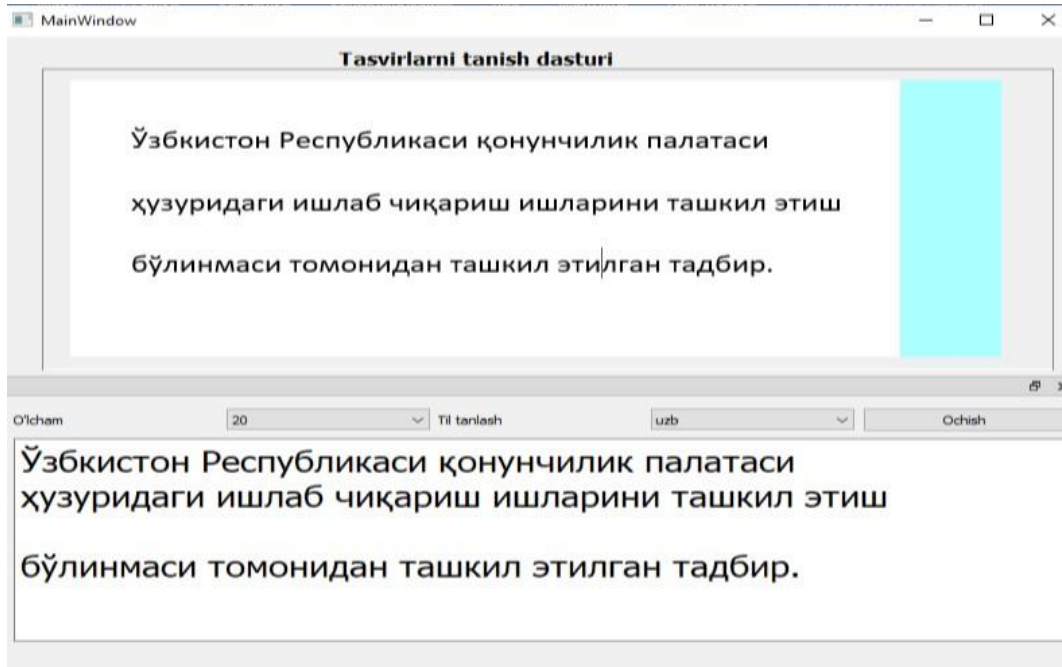
6-qadam: Matn bloklarini burchak nuqtasidan aniqlagandan so'ng, matn bloklarini tiklash uchun ushbu bloklarning ulanishi tekshiriladi.

G. Hisoblash natijalari. Bu sodda usul lekin aniqligi 90% dan yuqori, odatda o'rtacha 95%. Biroq, ushbu usul katta o'lchamdagi shriftlar uchun, burchaklari juda ko'p bo'lgan ba'zi bir maxsus rasmlar uchun unchalik samarali hisoblanmaydi. Keltirilgan muammolarga qaramay, u boshqa OCR vositalariga nisbatan tez (va parallellashtirilishi mumkin) va unchalik murakkab emas va kelajakda yanada yaxshilanishi mumkin. Shuningdek, ushbu usul paragraflar va chiziqlar kabi murakkabroq tartiblarni ajratib olishda juda samarali hisoblanadi.

Quyida o'zbek krill alifbosidagi matnni tanish natijasi keltirilgan. Tasvirlar kulrang darajadagi rasmlarga keltiriladi va ushbu rasmdan foydalaniladi. Shuning

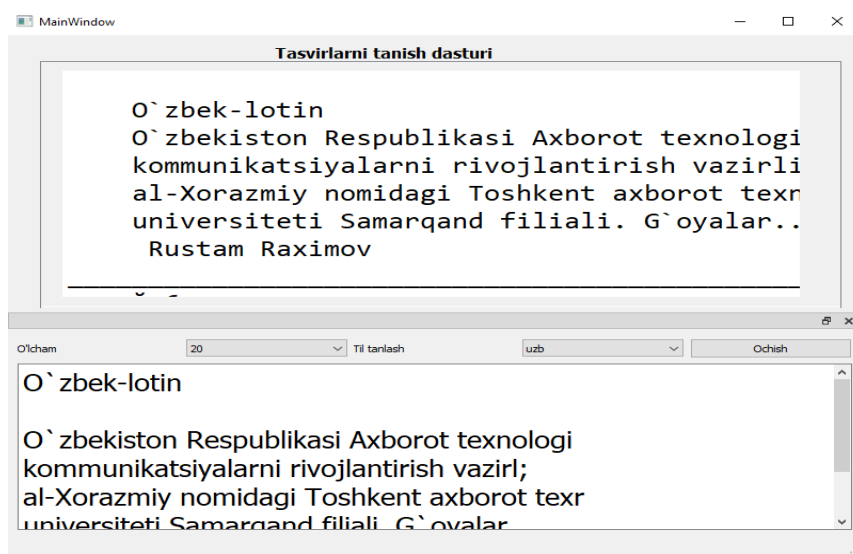
uchun tasvirni ikkilamchi ishlov berish talab qilinadi. Algoritm minimal taxminlarni keltirib chiqarmaydi.

4-rasm. O‘zbekcha (kril) natijasi.



Ushbu berilgan matndan qismlar hisoblab chiqiladi va burchaklar soniko‘proq bo‘lgan bloklarni matnli qismlarga, chegarasi kamroq bo‘lgan bloklaresa rasm yoki fon qismlarga ajratildi.

O‘zbek lotin va o‘zbek krilni aniqlaganimizdan so‘ng, konvertatsiya va tarjima sxemasi orqali jummalarni bir alifbodan boshqasiga aylantirish uchun konvertorni ishlab chiqiladi.

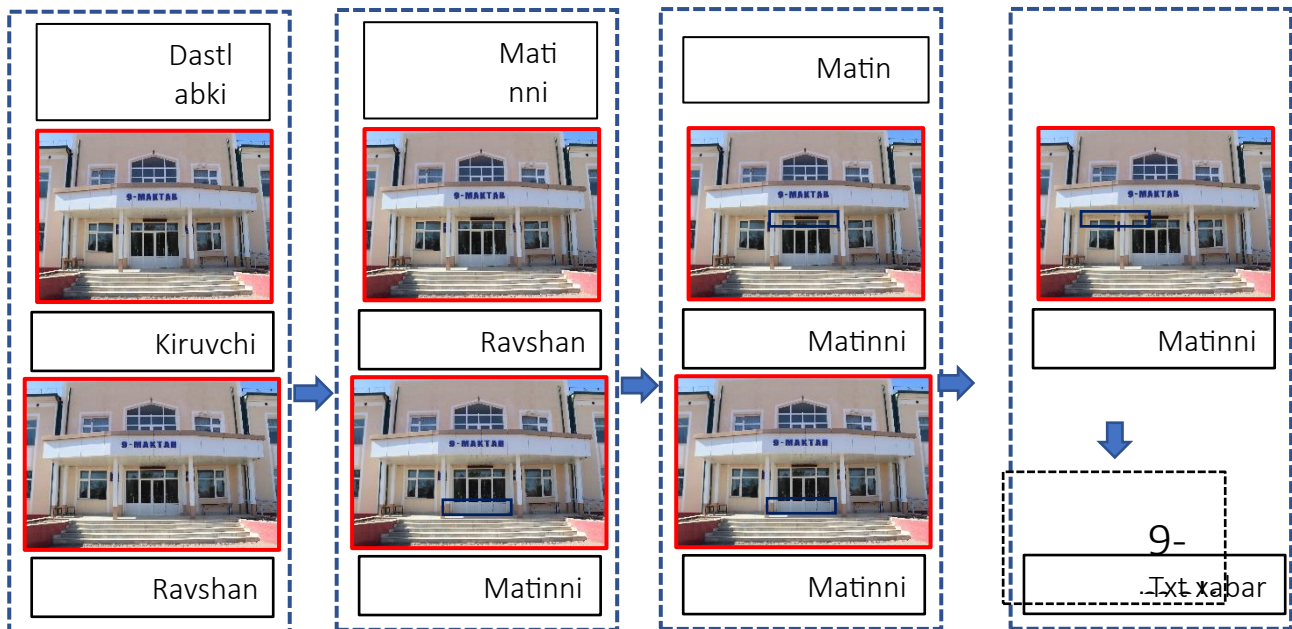


5-rasm. O‘zbekcha (lotin) natijasi

Matn bloklarini burchak nuqtasidan aniqlagandan so'ng, matn bloklarning (8 ta bog'langan qismlar) ulanishi tekshiriladi. Ushbu olingan natijalarda o'zbek lotin va o'zbek kril yozuvi natijalari keltirilgan. Qo'lda yozilgan hujjatlarda qator va so'zlarni segmentatsiyalashsamaradorligi keyingi ishlarda ko'rib chiqiladi. Qo'l yozma matni tanishda so'zlar va so'z birikmalari bo'sh joy bilan ajratilgan deb qaralgan.

Shuningdek, usul oldindan tayyorgarlikni talab qilmaydi, ya'nikim data set uchun alohida harflarni o'qitish yoki so'z shakllarini o'qitishni talab qilinmaydi. Lekin so'zlar va so'z birikmalari uchun data set tashkil etishda shu kitob manba uchun tashkil etiladi. Qancha ko'p qator o'qitilsa so'zlari olinsa tanish foizi oshadi.

Ushbu yondashuvda biz har qanday sifatli yo'naltirilgan rasmlardagi burchak nuqtalari orqali arzon narxlarda va parametrlar haqida ko'p ma'lumot ishlatmasdan aniq matn chiqarishni olish juda oson bo'lishi mumkinligini ko'rdik. Tasvirlardan matni olish uchun biz FAST algoritmgiga asoslangan juda oddiy usuldan foydalandik.



6-rasm. FAST algoritmgiga asoslangan natijasi

Ushbu usul juda tez, ko'p qirrali bo'lib, u yordamida turli xil tillarni, qo'l yozuvi va hattoki shovqin va xiralashgan tasvirlarni aniqlash mumkin. Ammo, bu uslub qo'l yozma va ba'zi bir maxsus rasmlar uchun ishlamayapti. Ushbu muammolarga qaramay, u boshqa OCR vositalariga nisbatan tez (va parallellashtirilishi mumkin) va unchalik murakkab emasva kelajakda yanada yaxshilanishi mumkin. Va nihoyat, ushbu usul paragraflar va chiziqlar kabi murakkabroq tartiblarni ajratib olishda juda samarali ko'rinadi.

Qo'lda yozilgan yozuvlarni taniy oladigan OCR tizimiga talab katta. Bu rasm uchun klaviatura yozish va shrift kodlashdan qochadi. Ushbu usul tasvir belgilarni

95% aniqlikda aniqlashda yordam beradi.

Nutqni tanib olish bugungi kunda eng muhim dasturlardan biridir. Taniqli bosma yoki qo'lda yozilgan OCR yozib olinishi va nutq chiqishi mumkin edi. Bu ko'zi ojizlarga ma'lumot yuborish va qabul qilishga yordam beradi.

FOYDALANILGAN ADABIYOTLAR RO'YXATI: (REFERENCES)

1. "Muhammad-al-Xorazmiy avlodlari" Ilmiy amaliy va axborot tahliliy jurnali №2 2022 yil Umarov M.
2. Suruchi G. Dedgaonkar, Anjali A. Chandavale, Ashok M. Sapkal, —Survey of Methods for Character Recognition // International Journal of Engineering and Innovative Technology (IJEIT), Volume 1, Issue 5, May 2012, ISSN: 2277-3754.
3. Shalin A. Chopra, Amit A. Ghadge, Onkar A. Padwal, Karan S. Punjabi, Prof. Gandhali S. Gurjar, —Optical Character Recognition // International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2014, pp. 4956-4958, ISSN (Online) : 2278-1021, ISSN (Print): 2319-5940.
4. Sarika Pansare, Dhanshree Joshi, —A Survey on Optical Character Recognition Techniques // International Journal of Science and Research (IJSR), Volume 3 Issue 12, December 2014, pp. 1247-1249, ISSN (Online): 2319-7064.
5. Sukhpreet Singh, —Optical Character Recognition Techniques: A Survey // Journal of Emerging Trends in Computing and Information Sciences, Vol. 4, No. 6 June 2013, pp. 545-550, ISSN 2079-8407.
6. Deepika Ghai, Neelu Jain, —Text Extraction from Document Images- A Review // International Journal of Computer Applications (0975 – 8887), Volume 84 – No 3, December 2013, pp. 40- 48.
7. Keechul Junga, Kwang In Kim, Anil K. Jain, —Text information extraction in images and video: a survey // Pattern Recognition, 37, pp. 977-997, 2004.
8. Line Eikvil, —Optical Character Recognition // Norsk Regnesentral, P.B. 114 Blindern, N-0314, December 1993.
9. Vikas Yadav, Nicolas Ragot, —Text extraction in document images: highlight on using corner points // in 2016 12th IAPR Workshop on Document Analysis Systems, pp. 281-286.