

АРАБ ГРАФИКАСИ ТАСВИРДАН ТАНИБ ОЛИШНИНГ АНАЛИТИК ТАҲЛИЛИ

М.Убайдуллаев,
mktiu93@gmail.com

Самарқанд давлат чет тиллар институти магистранти,
Х.А.Примова,
primova@samuit.uz

Муҳаммад ал-Хоразмий номидаги Тошкент ахборот технологиялари университети
Самарқанд филиали профессори

Аннотация. Араб графикаси услубида хилма-хиллик ва босма матнга нисбатан қўлёзма ёзуви сифатсизлиги уни машинада ўқитадиган матнга ўтказишда асосий муаммолар келтириб чиқармоқда. Ушбу мақолада араб тграфикасида ёзилган қўлёзма ва тошбосма асарлардаги матнни таниб олишдаги муаммолар кўриб чиқилади. Араб қўлёзмасида ҳарфларининг бири-бирига киришиб кетиши ҳамда устма-уст тушиши, лигатураларнинг жуда кўп ишлатилиши ва бошқа хусусиятлари сабабли матнни таниб олишда қийинчиликлар юзага келади. Шунинг учун араб графикасидаги матнни тасвирдан таниб олиш усулининг умумий асослари кўриб чиқилади.

Калит сўзлар: *Араб графикасидаги матнни таниб олиш, OCR, қўлёзма, тошбосма, тасвирдан матнни ўқиш, араб курсив ёзуви.*

Мамлакатимиз ижтимоий-сиёсий, иқтисодий-маиший, маънавий-диний каби соҳаларда янги ривожланиш босқичига ўтди. Илмий тадқиқотлар ривожини учун Президент Шавкат Мирзиёев фармонлари ва ҳукумат қарорлари методологик асос бўлиб хизмат қилмоқда. Жумладан, 2017 йил 24 майдаги “Қадимий ёзма манбаларни сақлаш, тадқиқ ва тарғиб қилиш тизимини янада такомиллаштириш чора-тадбирлари тўғрисида”ги ва бошқа қарорларда буюк аجدодларнинг асарлари ва уларнинг ислом дини ривожига қўшган ҳиссасини ўрганиш, ёш авлодда миллий ғурур ҳамда ифтихор туйғуларини тарбиялашдаги аҳамияти белгиланган [1, 2].

Жаҳонда қўлёзма матн тасвирига рақамли ишлов бериш, қўлёзма матн тасвирини босма матнга ўтказиш ҳамда қўлёзма матн тасвирини таниб олиш усул ва алгоритмларини такомиллаштириш, ишлаб чиқиш ва жорий қилишга катта эътибор қаратилмоқда. Босма матнларни таниб олиш ечилган муаммо деб ҳисобланган бўлсада, қўлёзма матни тасвирини таниб олиш ҳали ҳам муаммоли масала ҳисобланади [3].

Одамлар ёзувдан алоқа шакли сифатида фойдалана бошлагандан буён қоғоз энг устун ёзув воситаси ҳисобланади. Аммо замонавий электрон техника воситалари қоғознинг ўрнини эгаллаб бормоқда. Электрон воситалар жойини тежаш ва тез ишлаш имкониятлари билан оммалашиб бормоқда. Қоғознинг афзаллиги, унинг алоқа ва архивлаш жараёнида кенг қўлланилиши ва ҳозирги кунгача қоғозда ёзилган маълумотларнинг ниҳоятда кўплиги уларни автоматик тарзда ўқиш ва электрон шаклга солиш учун тезкор, самарали ва аниқ усулларни талаб этади.

Машинани инсон каби ўқитиш, яъни оптик белгиларни аниқлаш (OCR — optical character recognition) сўнгги 50 йилдан кўпроқ давр мобайнида кенг қамровли тадқиқотларнинг мавзусига айланди. Тасвирдан матнларни таниб олиш — бу қўлда ёзилган курсив матн ва машинада ёзилган босма матнни инсон каби ўқиш қобилиятини симуляция қилувчи андоза асосида ишловчи дастурий воситадир. Ҳозирги тизимлар одамлардан кўра тезроқ ўқий олиш имкониятига эга, аммо улар турли вариантлардаги матнларни ўқиш ва контексни тушуниш борасида етарли даражада ишончли эмас. Шунини айтиш мумкинки, инсон ўқиши ва машинанинг ўқиш қобилияти орасидаги тафовутни камайтириш учун катта ҳажмдаги қўшимча тадқиқотлар талаб этилади. OCR иловаларининг амалиётда зарурлиги ва OCRдаги муаммоларнинг қизиқарли томони бу борада изланишлар олиб боришга ва катта ютуқларга эришишга сабаб бўлди [6-7].

Ҳозирги кунда лотин алифбосидаги ёзувлар учун OCR тизимлари 99 фоиздан юқори таниб олиш имкониятига эга бўлиб, шахсий компьютерларда кенг фойдаланилади. Шу билан бирга, дастурий воситалар бозоридаги OCR тизимлари эндиликда турли ёзув услубларини (қўлда ёзилган, тошбосма ва босма шрифтлар) ҳамда хитой, япон, кирилл ва араб хатларидаги белгилар тўпламларини ўқий олади [3-5].

ا ب ت ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي

1-расм. Араб алифбоси ҳарфлари

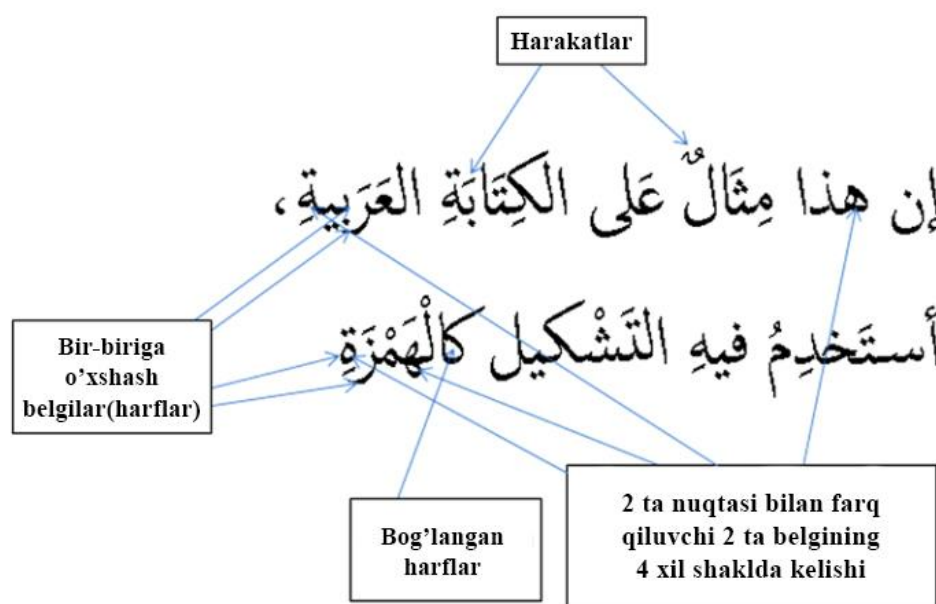
Арабча матн ўнгдан чапга қараб курсив шаклда ёзилади. Араб алифбоси 1-расмда кўрсатилганидек, 28 ҳарфдан ташкил топган бўлиб, ҳар бир ҳарф сўздаги жойлашган ўрнига кўра тўрттагача шаклга эга бўлиши мумкин: алоҳида кўриниши, сўзнинг бошида келадиган шакли, сўзнинг ўртасида ва охирида келадиган шаклларига эга (1-жадвал).

Сўздаги ҳарфлар бир-бирига боғланган ёки боғланмаган ҳолатда ва вертикал кўринишда ҳам келиши мумкин. Араб алифбоси ҳарфлари баландлиги ва кенглигига кўра турлича ҳажмда бўлади. Сўздаги ҳарфлар қисқа унлиларга (ҳаракатларга) эга бўлиши мумкин. Бу ҳаракатлар эса штрихлар кўринишида ёзилган бўлиб, ҳарфлар устига ёки остига қўйилади. Ҳарфларга қўйилган ҳар бир ҳаракат сўзнинг маъносини ўзгартиради. Ҳар бир ҳаракат эса рақамли матнга ўзгартирилаётганда ўзининг махсус коди (рақамига) эга ҳисобланади [3-5].

Asosiy ko‘rinishi	Kontextdagi shakllar		
	Boshida	O‘rtasida	Oxirida
ا			ا
ب	ب	ب	ب
ت	ت	ت	ت
ث	ث	ث	ث
ج	ج	ج	ج
ح	ح	ح	ح
خ	خ	خ	خ
د			د
ذ			ذ
ر			ر
ز			ز
س	س	س	س
ش	ش	ش	ش
ص	ص	ص	ص
ض	ض	ض	ض
ط	ط	ط	ط
ظ	ظ	ظ	ظ
ع	ع	ع	ع
غ	غ	غ	غ
ف	ف	ف	ف
ق	ق	ق	ق
ك	ك	ك	ك
ل	ل	ل	ل
م	م	م	م
ن	ن	ن	ن
ه	ه	ه	ه
و	و	و	و
ي	ي	ي	ي

1-жадвал. Араб алифбосидаги ҳарфларнинг шакллари

Араб графикасидаги матнни ўқийдиган мутахассислар контекстан маъно чиқариб, ҳаракатларни қўйиб ўқишга одатланган. 2-расмда араб ёзувидаги матннинг белгиларини таниб олишга оид айрим хусусиятлар кўрсатилган. Булар ҳарфларнинг асосий чизикда жойлашуви, бир-бирига ўхшайдиган ҳарфлар, ҳарфларнинг устига ёки остига қўйилган ҳаракатлар ва ҳарфнинг сўз бошида, ўртасида ва охирида келадиган шаклларида иборатдир [8].



2-расм. Арабча матндаги араб ҳарфларининг сўзда келган шакллари.

Қўлёзма матнни таниб олишдан кўра босма матнни таниб олиш жараёни осонроқ кечади, чунки арабча босма матнни таниб олишда бир мунча афзалликлар бор. Уларда ҳарфларнинг шакллари, сўзлар ва қаторлар орасидаги интерваллар ва бошқа хусусиятлар бир хилда бўлади. Бундан ташқари босма матн қаторларининг чизиғи қўлёзма матндаги қаторлар чизиғидан текисроқ бўлади. Шунинг учун босма матнни таниб олишда қатор чизиғига боғлиқ хусусиятлардан аниқроқ ва ишончлироқ фойдаланиш мумкин. Шунингдек, босма араб ёзуви матнини таниб олиш методларини

ишлаб чиқишда тадқиқотчилар бир қанча масалаларга эътибор бериши керак бўлади [5-7]:

- Курсив хусусиятга эга матн: араб графика матнининг босма шакли ҳам, қўлёзма шакли ҳам курсив тарзда ёзилади. Бу шуни англатадики, босма арабча матнни сегментациялаш (белгиларга ажратиш) лотин алифбосида ёзилган босма матнга нисбатан мураккаброқ жараён ҳисобланади. Бундан ташқари, араб ҳарфларининг усма-уст ёзилиши босма матнларда ҳам кенг қўлланилиб, сўздаги ҳарфларни сегментация қилишни қийинлаштирилади.

- Лигатуралар: араб графикасида ёзилган қўлёзма ва босма матнларда лигатуралар жуда кенг қўлланилади. Баъзи лигатуралар фойдаланилган ёзув турига боғлиқ бўлади. Лигатураларни сегментлаш қийин бўлиб, одатда ҳар бирини алоҳида белги сифатида кўриб чиқиш орқали моделлаштирилади.

- Ёзув шакллариининг кўплиги: Арабча босма матндаги бир қанча қулайликларга қарамай, араб графика ёзувидаги шакл(шрифт)ларнинг кўплиги матнни таниб олишда қийинчилик пайдо қилади. Ёзув турлари ичида ҳарфларнинг шакллари бир-биридан фарқ қилади. Ёзув шаклларидаги турлилик эвазига сўзлар орасидаги бўшлиқлар, ҳарфларнинг усма-уст тушиши, лигатураларнинг сони ва бошқа кўплаб фарқлар юзага келиши мумкин. Араб ёзувидаги матнни таниб олиш тизими иложи борида кўпроқ ёзув шаклларига ўргатилган бўлиши керак.

Хулоса. Араб ёзувининг курсив хусусияти, унинг қўшма ёзиш схемаси, лигатураларнинг кўплиги ва бошқа қийинчиликлар туфайли арабча матнни таниб олиш қийин вазифадир. Шунингдек, дунёдаги кўпгина халқлар араб алифбосига асосланган ёзувдан фойдаланади, буларга араб мамлакатларидан ташқари курд, форс ва уйғур каби бошқа халқлар киради. Мовароуннахр ўлкасидаги қўлёзма ва тошбосма асарлар ҳам асосан араб алифбосида ёзилган. Бу манбаларни электрон шаклга солиб, улардаги маълумотларни тасвирлардан ўқиб олишни автоматлаштириш муҳим вазифа ҳисобланади.

Фойдаланилган адабиётлар:

1. Ўзбекистон Республикаси Президентининг 2017 йил 24 майдаги ПҚ-2995-сонли “Қадимий ёзма манбаларни сақлаш, тадқиқ ва тарғиб қилиш тизимини янада такомиллаштириш чора-тадбирлари тўғрисида”ги Қарори.
2. Ўзбекистон Республикаси Президентининг 2019 йил 24 июндаги ПФ-519-сон "Ўзбекистон Республикаси Вазирлар Маҳкамаси ҳузуридаги Имом Бухорий халқаро илмий-тадқиқот марказини нинг кутубхона фондини хорижий адабиётлар билан бойитиш тўғрисида"ги Фармони.
3. Iskandarova S.N. To recognize the manuscript texts of Arabic letters in ancient Uzbek script//World scientific news an international scientific journal. – Polsha, 2019. Vol.115. –P.160-173
4. Имом Бухорий халқаро маркази Манбалар хазинаси Қўлёзмалар каталоги. Лойиҳа раҳбари Ш. Зиёдов. – Самарқанд. Имом Бухорий халқаро маркази нашриёти. – 2018. – 686 б.
5. Irfan Ahmed, Sabri A. Mahmoud, and Mohammed Tanvir Parvez. Printed Arabic Text Recognition. – 148 б. <https://www.researchgate.net/publication/278703169>.
6. Parvez, M.T., Mahmoud, S.A.: (2010) Polygonal approximation of digital planar curves through adaptive optimizations. Pattern Recognit. Lett. 31(13), 1997–2005
7. Pechwitz, M., Märgner, V.: (2003)HMM based approach for handwritten Arabic word recognition using the IFN/ENIT-database. In: Proc. Seventh International Conference on Document Analysis and Recognition (ICDAR), Edinburgh, Scotland, August 2003, pp. 890–894
8. http://www.bukhari.uz/wp-content/uploads/2018/08/PQ_2855.pdf