TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tarkvarateaduse instituut

Keir Volas 153012IAPM

OSTUTŠEKKIDE DIGITALISEERIMINE

Magistritöö

Juhendaja: Ermo Täks
Ph D

Tallinn 2017

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Keir Volas

08.05.2017

Annotatsioon

Käesolev lõputöö seisnes raamatupidamise pabertšekkide automatiseerimises. Töös selgitati, kui hästi on võimalik paberkandjal olevaid arveid viia digitaalsele kujule – kas seda on võimalik teha inimest kaasamata ja millise eelise annaks see manuaalse protsessi ees.

Selleks uuriti põhjalikult ärilist tausta, lahendusi ja korraldati küsitlus. Töös realiseeriti pabertšekkide digitaliseerimise prototüüp, milles töödeldi pilti, teostati optiline tekstituvastus, leiti tekstist andmed regulaaravaldiste ja valideerimiste abil ja koostati e-arve. Prototüübi tööd hinnati andmete leidmise, õigsuse ja ajakulu järgi.

Loodud lahendus suutis tuvastada täielikult ja õigesti 74% tšekkidest ning digitaliseerimise ajakulu oli keskmiselt 8 sekundit, mis annab ajavõiduks manuaalse protsessi ees 94%. Lisaks tõestati, et loodud lahendus aitab algdokumendi protsessi hõlbustada. Samuti leiti ideid, kuidas prototüüpi paremaks muuta ja võimalikke alternatiivseid kasutuskohti kogu lahendusele või osadele.

Töö tulemusena veenduti, et protsessi peaks jääma sisse manuaalne kontrollimise etapp. Sellest olenemata annaks antud lahendus juba praegusel hetkel märgatava ajavõidu ja hõlbustaks raamatupidajate tööd.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 50 leheküljel, 7 peatükki, 30 joonist, 1 tabelit.

Abstract

Digitalizing paper receipts

Bookkeeping has a lot of manual processes and they can be automated using technology. Entering paber receipts to bookkeeping software process was as time consuming and routine process. Aim of thesis was to find out how well it is possible to digitalize paper receipts – is it possible to do it without human interaction and how big advantage it will give compared with manual process.

Firstly, research of the business background was done, also were compared similar solutions and a survey was carried out between the bookkeepers. The main part was to build a prototype which is digitalizing paper receipts. Prototype consists following parts: preprocessing image, optical character recognition, info extraction from text using regular expression and validation, e-invoice creation. Finally, prototype was validated against 90 pictures of receipts and data finding, validity, time consumption were evaluated.

Created solution was able to digitalize fully 74% of receipts and time consumption was 8 seconds which is 94% less than manual process. Additionaly, it was proved that solution helps to preserve original accounting document more easily. Also new ideas were suggested how to improve prototype and where else can it be used than in this bookkeeping process.

As a result of this thesis, it was convinced that paber receipt entering process can be improved. Still, data validity check by human should stay in this. Despite of this, solution would give a significant time saving in process and would make bookkeepers life easier.

The thesis is in Estonian and contains 49 pages of text, 7 chapters, 30 figures, 1 table.

Lühendite ja mõistete sõnastik

BASE64	Kodeerimismeetod, mis teisendab binaarandmeid ASCII tekstiks ja vastupidi.
FREJ	*Fuzzy Regular Expressions for Java*, hajusad regulaaravaldised Java jaoks.
ITIL	Eesti Infotehnoloogia ja Telekommunikatsiooni Liit
KMKR	Käibemaksukohustuslase number
OCR	*Optical Character Recognition*, optiline tekstituvastus.
PDF	*Portable Document Format*, porditav dokumendiformaat.
XML	*Extensible Markup Language*, laiendatav märgistuskeel.
XSD	*XML Schema Definition*, XML skeemi definitsioon.

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Tehnoloogia areng asendab aina enam manuaalseid tööprotsesse pea igas valdkonnas. Finantssektor ja eelkõige pangad on muutumas aina rohkem infotehnoloogia ettevõteteks. Samas on raamatupidamise valdkond püsinud küllaltki sarnasena. Töö autor näeb oma magistriõppes omandatud teadmiste põhjal mitmeid võimalusi hõlbustada raamatupidajate tööd, kui digitaliseerida manuaalseid protsesse infotehnoloogia poolt pakutavate võimalustega.

1.1 Taust ja probleem

Raamatupidamises kulutatakse palju aega andmete dubleerimisel erinevatesse süsteemidesse, mis eelkõige hõlmab enda all manuaalset andmesisestust. Ostuarvete andmesisestus on üks tavapärastest raamatupidamise protsessidest. Ostuarveteks on suurel hulgal paberil olevad kassatšekid ja arved, mida tuleb sisestada raamatupidamistarkvaras manuaalselt. Manuaalne andmesisestus on vägagi ajakulukas ning sellise tööprotsessi hõlbustamine infotehnoloogia vahenditega on vägagi väljakutsuv.

1.2 Ülesande püstitus

Teadustöös eesmärgiks on automatiseerida raamatupidamises ressursikulukaid tööprotsesse. Raamatupidamise tööprotsessi on võimalik oluliselt kiirendada ja parendada, kui automatiseerida arvete sisestamist. Täpsemalt keskendutakse pildikujul olevate arvete digitaliseerimisele, mis omakorda hoiaks kokku raamatupidajate rutiinset ja käsitsi tehtavat tööd.

Teadustöös seatud uurimusküsimus ja toetavad küsimused on järgnevad:

1. Kas raamatupidamise tööprotsessi on võimalik oluliselt kiirendada ja parendada, kui automatiseerida arvete sisestamise etappi?

1.1. Kas paberkandjal olevate tšekkide viimine digitaalsele kujule annab märgava eelise võrreldes manuaalse sisestusprotsessiga?

1.2. Mis määral saab paberkandjal olevat tšekki viia digitaalsele kujule ilma inimest kaasamata (kas loodav lahendus suudaks väljastada sobiva digitaalse arve või peab inimene seda kontrollima ja täiendama)?

1.3 Metoodika

Teadustöös uuritakse ärilist tausta, milleks on raamatupidamise valdkonnas ostuarvete sisestuse protsess. Täpsemalt selgitatakse välja üldlevinud standardid ning korraldatakse uuring raamatupidamist pakkuvate ettevõtete raamatupidajate vahel, selgitamaks välja tööprotsessid ja paberarvete mahud. Lisaks uuritakse relevantseid teadusartikleid ja võrreldakse seotud töid.

Eelmainitu põhjal kujundatakse loodava prototüübi ülesehitus. Prototüüp teostaks pildifaili põhjal e-arve genereerimist, kasutades selleks välja uuritud parimaid variante ja metoodikaid. Samuti valideeritakse prototüübi tööd, mille alusel hinnatakse lahenduse kasutatavust käesoleva raamatupidamise protsessi parendamiseks.

1.4 Ülevaade tööst

Magistritöö esimene peatükk juhatab töö sisse, annab aimu probleemist, eesmärkidest ja töö teostamise metoodikast.

Teises peatükis on toodud põhjalik taustauuring ärilise probleemi kohta ning on toodud võrdlus olemasolevate seotud lahenduste kohta. Samuti viidi raamatupidajate vahel läbi uuring, mille tulemused on toodud teise peatüki lõpus.

Kolmas peatükk keskendub teoreetilisele taustale, mille esimene pool selgitab ärilist raamatupidamisalast infot. Järgmises osas on toodud probleemi lahendamiseks kasutatavad infotehnoloogia suunad. Viimases osas on analüüsitud kõige sarnasemaid tehtud töid ning toodud välja nende puudused.

Neljas ja viies peatükk on keskendunud loodud lahenduse tutvustamisele. Neljandas peatükis antakse ülevaade prototüübi jaotamisest ja nende osade realiseerimisest. Nendeks osadeks on pildi töötlemine, optiline tekstituvastus, väljade märgistamine ja e-

arve genereerimine. Viies peatükk keskendub prototüübi valideerimisele, kus on kirjeldatud prototüübi testimine vastavalt seatud eesmärkidele ning on välja toodud tulemused.

Kuuendas peatükis analüüsitakse töö käigus teada saadud infot ja prototüübi tulemusi ning antakse hinnang lahenduse toimimise, parendamise ideede ja edasiste kasutuskohtade kohta.

Töö kokkuvõte on toodud seitsmendas peatükis, kus võetakse lühidalt käesolev töö kokku ning antakse ülevaade seatud eesmärkide vastavusest.

2 Probleemi analüüs

Käesolevas töös lahendatav probleem seisneb andmesisestuse protsessi parendamises ning tuleneb raamatupidamise valdkonnast. Raamatupidamine tegeleb ettevõtte majandustehingute arvestuse pidamisega. Eestis on 2016. aasta seisuga 154781 majanduslikult aktiivset ettevõtet [1]. Neil kõigil lasub raamatupidamise kohustus [2].

2.1 Probleemi teke

Kulukas manuaalne dokumentide sisestus on alati probleemiks olnud paljudes valdkondades. Raamatupidamises on palju erinevate dokumentide sisestamist ja ajapikku on seda parandatud erinevate raamatupidamistarkvarade ja nende omavahelise suhtluse teel. Seda probleemi on samuti mingil määral aidanud lahendada e-arvete kasutuselevõtt. See pole aga aidanud kõikide dokumentide puhul ja siiani on paberkandjal olevad dokumendid kasutusel kuludokumentide haldamiseks. Tänu tehnoloogia arengule on võimalik seda probleemi optilise tekstituvastuse ja semantika abil lahendada. Eelmainitud tehnoloogilised lahendused on piisavalt heaks saanud ning Eestis on kasutusele võetud standardiseeritud e-arve formaat, mida toetavad raamatupidamistarkvarad. Seetõttu on tekkinud võimalik lahendus käesolevale probleemile.

2.2 Olemasolevad lahendused

Pikalt arutletud ja plaanitud muuta paberarved elektrooniliseks. 2010a teavitas Eesti Infotehnoloogia ja Telekommunikatsiooni Liit (ITL) e-kviitungi lahenduse välja töötamisest [3]. 2011. aasta lõpus pidavat e-kviitungi lahendus valmima, mis asendaks pabertšekid ainult kaardimakse korral [4]. Lahendus ei valminud nii aastal 2011 kui ka järgnevalt lubatud 2012. aasta jooksul. Vahepeal jõudsid suuremad kaupluste ketid ette ning võtsid kasutusele oma e-tseki, mille eesmärgiks oli pigem olemasolevatele kasutajatele pakkuda iseteeninduses kulude jälgimist ja tšekkide säilitamist garantiiperioodiks [5].

2016. aastal teavitati järjekordsest e-kviitungi projekti avamisest, mille eestvedajaks on Omniva [6]. Sama aasta lõpus pidi portaal avatama lõpptarbijatele, kuid 2017. aasta algul oli projekt siiani pilootfaasis [7].

Käesoleva töö raames plaanitavat tšeki automaatset ja kohest sisestuse protsessi lahendust teadaolevalt turul hetkel ei ole. Üks ettevõtte on maininud sellise lahenduse plaanist. Seega tuuakse järgnevalt välja teemaga seotud äriliste protsesside pakkujad, nende teenuste kirjeldused ja erinevused.

2.2.1 E-kviitung

E-kviitungi lahendus hakkab pakkuma portaali, mis võimaldab lõpptarbijatel hallata veebikeskkonnas oma kviitungeid ning nendega kaasnevaid dokumente (garantiikirjad, tootejuhendid). Kaupmeestele pakuks E-kviitungi lahendus loobuda paberkviitungitest ning võimalus suunata kviitungid otse raamatupidamisse. Samuti kasutatakse laiendatud e-arve standardit. [8]

Protsess näeks välja järgmine:

1. Kaupmehe äritarkvara on integreeritud e-kviitungi lahendusega
2. Järgnevad ostuprotsessi osad seovad tehingu isikuga:
    a. Makse pangakaardi või mobiiliga
    b. Kliendikaardi kasutamine
3. Digitaalne kviitung jõuab E-kviitungi portaali

Kogu selle lahenduse realiseerimine tähendaks seda, et kogu süsteemi peaks muutma, kõik kaupmehed peaksid oma äritarkvarad selle lahendusega integreerima. Lisaks peaksid kõik tarbijad kasutama kliendikaarti või maksma pangakaardi või mobiiliga. Alles siis oleks võimalik täielikult pabertšekkidest vabaneda.

Selliste uute protsesside kasutuselevõtt ja laialdased integratsioonid võtavad väga kaua aega ning selleks kõiki kaupmehi kohustuda on keeruline. E-arvetele üleminek avalikus sektoris viibib siiani, mille sisuks oli e-arvete vastuvõtmine ainult 50% asutuste seas 2016. aasta alguseks [9]. Seega pabertšekid ei saa lähima aastate jooksul täielikult kaduda.

2.2.2 Tsekk.ee

Tsekk.ee on kuludokumentide haldamise vahend, mis aitab kokku koguda kviitungeid, neid sorteerida, arhiveerida, digitaliseerida ja raamatupidamises kajastada. See on mõeldud paberkandjal olevate kuludokumentide haldamiseks. Sellise lahenduse loomist alustati juba 2014. aastast. 2016. aasta lõpus oli süsteem poolikult kasutatav. Kõikidele kasutajatele avati süsteem uuel kujul 2017.a märtsis. Tsekk.ee lahendusel on integratsioonid peamiste kasutatavate majandustarkvaradega. [10][11]

Protsess näeb välja järgmine:

1. Pildista tšekki
2. Lae pilt üles veebilehel või mobiilirakenduses
3. Sisesta tšeki peal olevad andmed veebilehel (automaatset digitaliseerimist ei toimu) või oota nende poolt andmesisestust
4. Ekspordi digitaalne kuludokument järgmistel võimalikel viisidel:
    a. Saada integreeritud süsteemi (võimalik järgnevate süsteemide puhul: Fitek, Erply, Merit Aktiva ja Directo)
    b. Ekspordi laiendatud e-arve

Antud lahendusel ei toimi kohene andmete digitaliseerimine. Seal kasutatakse andmete digitaliseerimiseks manuaalset viisi, mida ostetakse sisse lepinguliste koostööpartnerite käest. Selline manuaalne sisestamine on ikkagi aja- ja rahakulukas ehk Tsekk.ee'd kasutades on viidud see töö raamatupidajalt teiste inimeste kätte. 2016. aasta septembris on mainitud, et Tsekk.ee'l on valmimisel programm, mis kviitungite andmed automaatselt sisse loeb [12]. 2017. aasta aprilliks ei ole see veel valminud ning kasutajatel pilti üles laadides automaatset andmete digitaliseerimist ei toimu [10].

2.2.3 Envoice

Envoice on arvete ja dokumentide digiteerimise ja haldamise tarkvara, mis on loodud samuti lähiaastatel ning juurutamine on täies hoos. Ettevõte suund on muuta raamatupidamine paberivabaks. Loodud e-arved on aidanud seda saavutada ning

Envoice pakub laia valikut erinevaid teenused, mille all on kuluhaldus, eelarve jälgimine ja automatiseeritav müügiarvete koostamine. [13]

Selliste teenustega turule tulemine tõestab samuti käesoleva lõputöö aktuaalsust, et automatiseerida raamatupidajate tööprotsesse. Envoice lahendusel puudub automaatne tšekkide digitaliseerimine, mis on just käesoleva töö plaanitav osa.

2.3 Lahenduse huvipooled

Lahenduse huvipoolteks oleks eelkõige:

- Raamatupidamisteenust pakkuvad ettevõtted (raamatupidajatel kulub palju väärtuslikku aega manuaalsele andmesisestusele)
- Suured ettevõtted, kellel tekib palju kuludokumente

Käesoleva töö tõestatud lahendusest võib olla huvitatud ka eelmainitud kuludokumentide haldamist pakkuvad ettevõtted, sest ühel neist on käesoleval hetkel käimas sama probleemi lahendamine ning see annaks nende teenusele väga suure eelise.

2.4 Uuring

Käesoleva töö raames viidi raamatupidajate vahel läbi uuring, saamaks teada töö jaoks vajalikku infot seoses paberkandjal olevate kuludokumentidega. Lisas 1 on toodud välja küsitluse ankeet. Küsitluses osales 6 raamatupidajat, kes töötasid erinevates raamatupidamisteenust pakutavates ettevõtetes. Igal raamatupidaja halduses oli rohkem kui üks keskmises suuruses ettevõtet. Küsitluses osalenute arvu ei aetud liiga suureks, sest statistika asemel oli eesmärk pigem saada kindlust probleemi aktuaalsusele juba väikese hulga raamatupidajate seas ning põhieesmärk oli saada ärilist tausta, mis on raamatupidajatel ühine.

Samuti olid raamatupidajad nõus jagama infot ning nendega suheldi aktiivselt ka peale küsitlust, et töös lahendatud prototüüp oleks äriliselt korrektne.

2.4.1 Pabertšekkide osakaal

Uuringus küsiti iganädalaselt sisestavate paberkandjal olevate tšekkide arvu ja nendele kuluvat aega. Selgus, et tšekkide hulk ja sisestuse aeg varieerub üsna palju. Üks

raamatupidaja sisestas pabertšekke 10 tükki nädalas ja teine 80 tükki nädalas. Ajakulu oli 30 sekundist kuni 4 minutini. Lähemalt põhjust uurides saadi teada, et see oleneb suurel määral ettevõtete suurusest ja tegevusalast. Suurematel ettevõtetel on palju suuremal määral paberkandjal olevaid kuludokumente.

Keskmisena sisestatakse igakuiselt 148 paberkandjal tšekki. Ühele sisestusele kuluv keskmine aeg on 2,1 minutit.

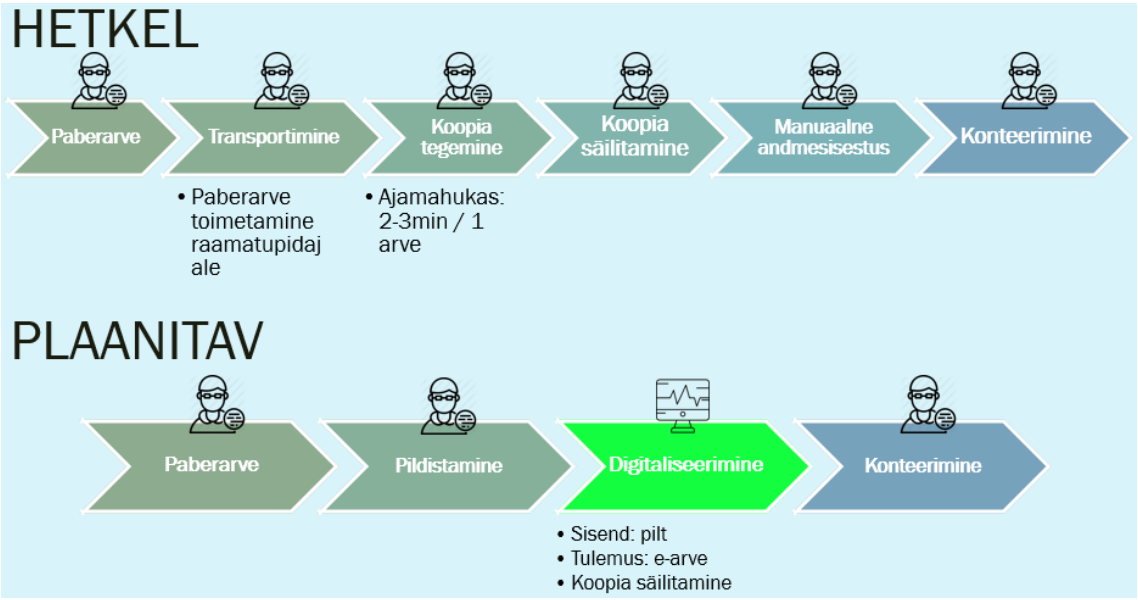
Lisaks küsiti e-arvete osakaalu kogu kuludokumentide seast. Enamikul raamatupidajatel ei olnud kuludokumentide seas ühtegi e-arvet, maksimaalselt oli ühel raamatupidajal e-arvete osakaal 30%.

2.4.2 Pabertšekkide digitaliseerimine

Kaks raamatupidajat kuuest ei olnud kuulnud midagi paberkandjal olevate kuludokumentide sisestuse automatiseerimisest. Teised neli olid kuulnud, aga kasutanud ei olnud nendest keegi. Lisaks avaldasid kõik kuus uuringus osalenud raamatupidajat arvamust ja andsid soovitusi pabertšekkide digitaliseerimise kohta. Enamikule meeldis mõte väga, sest see lihtsustaks nende teed suurel määral.

2.4.3 Protsessi kirjeldus

Samuti uuriti raamatupidajatelt tegelikku paberarvete andmesisestuse töökäiku. Teada saadud paberarvete üldine haldamise protsess on toodud joonisel 1.



Joonis 1. Paberarvete sisestuse protsessivaade (hetkel ja plaanitav) .

Jooniselt 1 on hästi näha mitmeid erinevaid manuaalseid protsesse. Nende hulgas on paberarve transportimist raamatupidaja juurde, mille käigus toob ettevõtja dokumendid ise raamatupidaja juurde ja mõnedel juhtudel käib raamatupidaja ise ettevõtja juures neil järgi. Seejärel tehakse tšekkidest koopiad ning nende säilitamiseks peab need korrektselt õiges kohas arhiveerima. Seejärel toimub andmesisestus raamatupidamisprogrammi, mille järel konteeritakse kanne. Konteerimise all valitakse õiged kategooriad ja kontod ning kinnitatakse kanne.

Joonisel 1 on rohelisena toodud käesoleva töö loodav lahend. Nagu näha, siis annaks see etapiliselt suurt kokkuhoidu.

2.5 Aktuaalsus

Teemapunkti 2.2 all mainitud olemasolevate lahenduste järgi on käesoleva töö probleem vägagi aktuaalne. Tsekk.ee tegeleb sama probleemiga ning nende kuludokumentide haldusvahend on saanud Eesti majandustarkvarade ja raamatupidamise ettevõtete seas populaarseks. [14]

Algselt väikse mahuga tundunud probleemi lahendamise mõju oleks päris suur. Uuringust selgus, et raamatupidajad hindaksid sellist lahendust väga. Keskmist pabertšekkide arvu ja sisestuse ajakulu arvesse võttes kulutab iga raamatupidaja igakuiselt üle 5 tunni intensiivset tööaega, mis on ligi üks tööpäev. Statistikaameti 2015. aasta uuringu järgi on Eestis raamatupidaja ametikohal töötavate inimeste arv 18000. Keskmine raamatupidamisteenuse tunnihind on 30 eurot. Seega teeks see Eestis ettevõtetele rahaliselt säästmist 2,7 miljonit eurot kuus. [15]

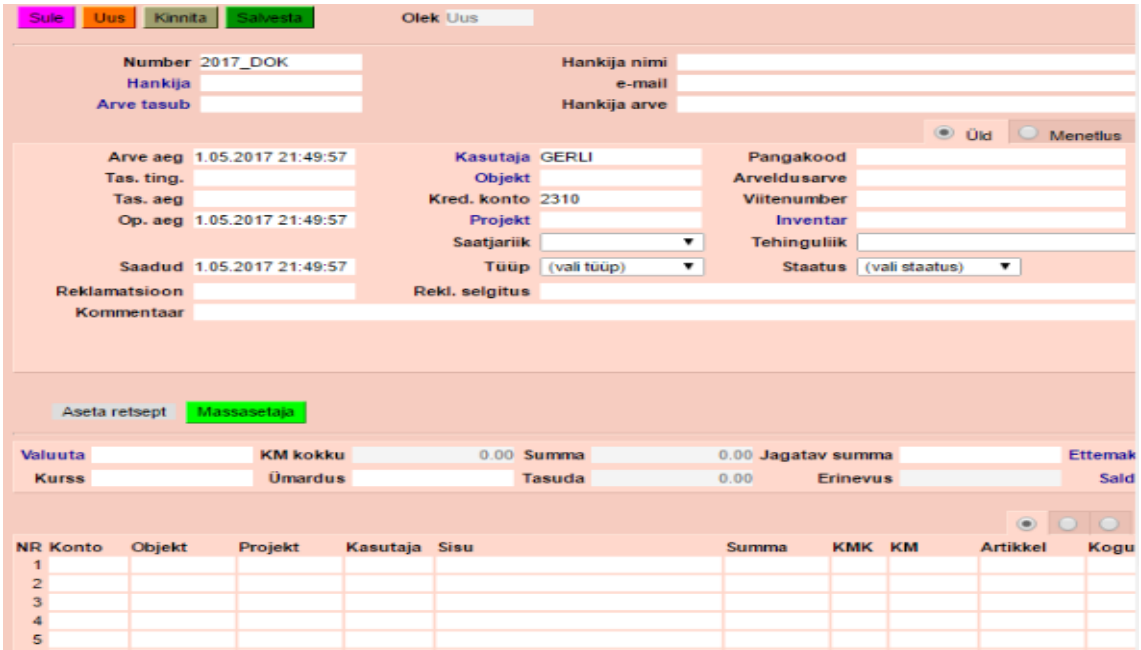
3 Teoreetilised lähtekohad

3.1 Raamatupidamisalane taust

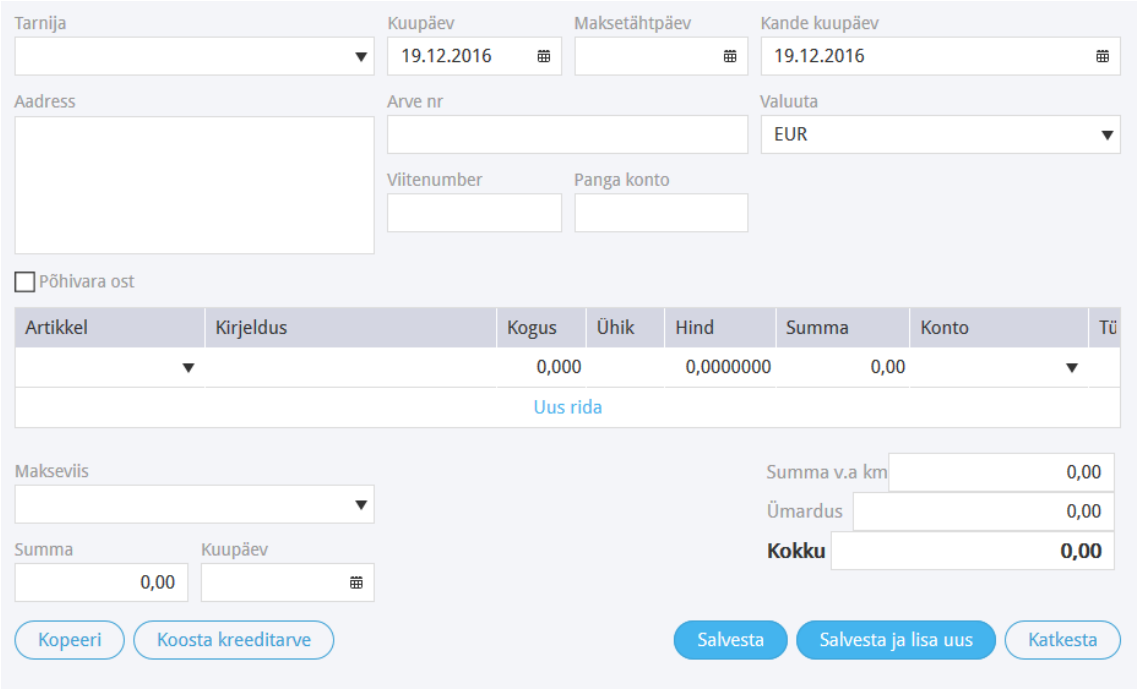
Raamatupidamise seaduse §6 kohustab majandustehinguid dokumenteerima ja kirjendama. Ostutšekk on majandustehingut tõendav algdokument, mis on aluseks raamatupidamiskirjendile. Raamatupidamiskirjendid sisestatakse raamatupidamistarkvarasse. Vastavalt eelmainitud seadusele peab kirjend sisaldama järgmisi andmeid:

- majandustehingu kuupäev
- kirjendi identifitseerimistunnus, näiteks number või numbri ja tähe kombinatsioon
- debiteeritavad ja krediteeritavad kontod ja vastavad summad
- viide kirjendi aluseks olevale alg- või koonddokumendile. [16]

Joonisel 2 on toodud raamatupidamistarkvara Directo ostuarve sisestuse vaade ja joonisel 3 on Merit Aktiva sisestuse vaade. Joonistelt selgub, et andmesisestuse pilt on üsna kirju, sest välju on vormil palju.



Joonis 2. Ostuarve sisestuse vorm programmis Directo.



Joonis 3. Kuludokumendi sisestuse vorm programmis Merit Aktiva.

3.1.1 E-arve

E-arve on masinloetav arve, mis on koostatud ühtse standardi alusel. Ühtne standard võimaldab e-arveid liigutada erinevate süsteemide vahel, et vältida andmete topelt sisestamist [17]. Ostuarvete digitaalne üldformaat eksisteerib XML-kujul, mida toetavad enamik raamatupidamistarkvarasid [18].

E-arved võimaldavad otse ostja panka saatmist ja tasumise infot. Kuludokumentide puhul ei ole sellist infot vaja esitada, sest E-arve formaati kasutataks raamatupidamistarkvarasse importimiseks ja algdokumendi säilitamiseks.

Järgnevalt toodud ainult e-arve kohustuslikud elemendid:

- *Header* – e-arve failipõhine info
    - *Date* – faili genereerimise kuupäev
    - *FieldId* – faili unikaalne identifikaator
    - *Version* – kasutatava standardi versioon
- *Invoice* – ühe arve info (neid elemente saab olla rohkem kui 1)

- *InvoiceId* – arve unikaalne identifikaator
- *RegNumber* – arve vastuvõtja registri- või isikukood
- *SellerRegNumber* – müüja registrikood
- *InvoiceParties* – arvega seotud osapoolte info
    - *SellerParty* – müüja andmed (*SellerPartyRecord*)
    - *BuyerParty* – ostja andmed (*BillPartyRecord*)
- *InvoiceInformation* – arve põhiandmete info
    - *Type* – arve tüüp
    - *DocumentName* – dokumendi nimi
    - *InvoiceNumber* – arve number
    - *InvoiceDate* – arve kuupäev
- *InvoiceSumGroup* – arvete summa info
    - *TotalSum* – arve summa kokku
- *InvoiceItem* – arve teenuste/kaupade info
    - *InvoiceItemGroup* – arve ridade grupp (võib olla rohkem kui 1)
    - *ItemEntry* – kirjeldab ühte rida arvel (võib olla rohkem kui 1)
        - *Description* – kauba/teenuse nimi või kirjeldus
- *PaymentInfo* – maksekorralduse genereerimiseks vajalik info
    - *Currency* – kolmetäheline valuuta kood
    - *Payable* – arve maksmise indikaator
    - *PaymentTotalSum* – maksmisele kuuluv summa

        - *PayerName* – ostja nimi
        - *PaymentId* – arve number
        - *PayToAccount* – müüja makse laekumiskonto
        - *PayToName* – müüja nimi

- *Footer* – e-arvega edastatud arvete hulga ja kogusumma info
    - *TotalNumberInvoices* – arvete arv failis (*Invoice* elemendid)
    - *TotalAmount* – elementide *PaymentTotalSum* kogusumm

Käesoleva töö raames tasub mainida mittekohustuslikku välja *AttachmentFile,* mis on mõeldud arve manuse hoidmiseks (*Invoice* alamelement). [19]

3.2 Tehniline taust

3.2.1 Optiline tekstituvastus

Optilise tekstituvastuse kohta leidub küllaltki palju infot ja eksisteerivat tarkvara. Nende suutlikkus ja täpsus varieerub väga suurel määral. Tavalisemate tekstide kohta on näitajad head ning optiline tekstituvastus on end tõestanud. [20]

Samuti on optilisel tekstituvastusel ka probleeme – tähtede müra on alles ja raske tähti õigesti tõlgendada. [21]

Ajapikku on aga tehnoloogia sel alal päris heaks läinud ning saab seda laialdasemalt ära kasutada. Optilise tekstituvastuse pakutud lahendusena on jagada esialgne sisu omaette väiksemateks osadeks. Esmalt tuleks eristada suuremad vahed ja tekstide vahelised eristused. Seejärel saaks neid osasid kergemini analüüsida – see osa peaks olema eraldiseisev (algus ja lõpp, laused või fraasid). Kui see osa osutub poolikuks, tuleks ühendada ta järgmise osaga kokku. Sealjuures võib välja tulla ka veergude eristamine. Eesmärk oleks leida terviklikke osasid suuremast dokumendist. [22]

3.2.2 Infootsing, väljade märgistamine

Semantilise osa, mis hõlmab tekstist aru saamist ja tõlgendamist vajalikule formaadile, kohta leidub samuti infot. Eelkõige põhineb see kontekstil. Tavaliselt kasutatakse semantikat peale optilist tekstituvastust, et saadud tekstist aru saada ja edasi kasutada. [23]

Käesolevas töös olev tekst ehk tšeki sisu on vägagi konkreetne ja piiratud kontekst, mille sisu põhineb pigem objekt-väärtus skeemil ning koosneb lühenditest ja koodidest. Piiratud kontekstiga teksti ja väärtuseid omades ei pruugi anda semantilise osa juurde lisamine võitu. Samuti laialdaselt kasutatud masinõppe meetodid ei ole soovitatud väga kitsa teemade peal, sest need toimivad paremini suuremate tekstidega. [24] Piiratud konteksti puhul näitavad siiani head tulemust reeglipõhised süsteemid. Tähtis on teada ärilist tausta ning reeglid luua, mis on ka tavaliselt raskeimaks osaks. Seda eelkõige realiseerija vaates, sest arendaja tunneb infotehnoloogia valdkonda hästi, kuid kõike muud süvitsi teada ei ole võimalik. Omistamise reeglitega on üldjuhul lihtsam ja kontekst piiritletum. [25][26]

Käesolevas lahenduses on samuti vaja omistamise reegleid, sest teada on konkreetsed vajatavad väärtused ja need on vaja tekstist leida. Seega on kontekst teada, mingil määral reeglid samuti, mida täiendatakse koos raamatupidajatega ja valideerimise teel.

3.3 Seotud tööd

Raamatupidamise paberdokumentide digitaliseerimist on püütud teha juba pikka aega – viimaseid aastakümneid. Küll on proovitud manuaalsete vigade auditeerimise protsesse ja süsteeme rakendada kui ka elektrooniliste arvete süsteeme juurutada. [27] [28]

Sellest hoolimata on siiani väga suur osa arvetest paber või tavalises PDF-kujul, mis tuleks sisestada manuaalselt raamatupidamistarkvarasse. Selline manuaalne sisestus on vägagi ajakulukas ja tekitab inimlikke vigu. Vead raamatupidamisdokumentides tekitavad probleeme ning veel suuremat ajakulu. Selliseid probleeme on üritatud veidi lahendada või leevendada, näiteks kasutades arve sisestuse kontrolli süsteemi. Esmane, täiesti suurte ja ebaloogilise info avastamine on püütud realiseerida automaatselt, kuid on jäätud süsteemi sisse teise inimese poolt manuaalse kontroll. [29]

Samuti on toodud häid näiteid protsessi ja lahenduse poolelt aitamaks digitaliseerida tööprotsesse, kuid need on liialt seotud täiesti uue süsteemi loomisele [30]. Selliste täiesti uute süsteemide loomine ja juurutamine on liialt keerukas. Tegelikult peaks keskenduma rohkem üldlevinud protsessidele ja leidma nende ühisosa, mida oleks võimalik paremaks muuta digitaliseerimisega.

Üldlevinud raamatupidamise protsessides on siiani alles paberarved ja nende manuaalne sisestamine, lisaks on kasutusel raamatupidamistarkvarad, mis toetavad digitaalse arve üldformaati [18]. Seetõttu oleks hea võimalus just seda tööprotsessi efektiivsemaks muuta, sealjuures kogu süsteemis suuri muudatusi tegemata.

Järgnevalt on toodud kõige sarnasemate lahenduste kirjeldus ja puudused.

3.3.1 Uurimus: pildipõhine arve tasumine

Sarnaseid lahendusi on välja toodud arvete maksmise protsessis. Protsess on analoogne: arve saamisel skaneeritakse dokument, genereeritakse pilt, skaneerimise seade teostab optilise tekstituvastuse ja eraldab maksmiseks vajaliku info, vajalikud andmed (skaneeritud pilt, arvelt tekst-tuvastatud andmed, muu makse info), saadud andmete põhjal tasutakse arve. [31]

Antud artiklis on toodud välja mõistlik võimalik äriline protsess, mis annab kinnitust ärilises vaates selliseid asju teostada. Kahjuks on antud töös praktiline teostus ja protsessi toimimise tõestus puudu.

3.3.2 Uurimus: automaatne arvete haldamine masinõppe ja OCR'iga

Rootsis uuriti samalaadset probleemi – automaatne arvete haldamise protsess kasutades masinõpet ja optilist tekstituvastust (OCR). Töös katsetati kahte erinevate optilise tekstituvastuse mootorit (Tesseract ja OCRopus). Tulemustena leiti, et ainuüksi optilise tekstituvastuse mootoritega ei olnud võimalik arveid kasutatavaks infoks tõlgendada, sest tulemuseks oli sisutühi tekst. Arvati, et optilisele tekstituvastusele tuleks lisada erinevate mallide järgi tuvastus. Prototüübis leiti, et masinõppe integreerimine võiks oluliselt aidata andmete korrektsust hinnata. [32]

Antud töös jäigi puudu veidi semantilisest osast, mida tulemustes ka mainiti (ilma tõlgendamise mallideta ei olnud optilise tekstituvastusega võimalik arveid tõlgendada). See tõestab veel kord, et äratundmise osa on tähtis ning koostöös optilise

tekstituvastusega tuleb leida just vajalikus formaadis info. Selleks võiks uurida arvete üldlevinud ülesehitust ning pookida see külge.

3.3.3 Uurimus: automaatne paberarvete lugemine ja tõlgendamine

2016. aasta mais on tehtud töö paberarvete automaatse lugemise ja tõlgendamise kohta [33].

Töös tuuakse välja erinevaid olemasolevaid vahendeid. Näiteks on Rootsis eraisikutele kasutamiseks Swedbank'i mobiilirakendus, mis võimaldab arve maksmiseks vajalikud andmed kaamera kaudu paberarvelt valima (peab ise pildil näitama, kus õiged andmed asuvad). Töös tuuakse välja ka erinevad optilise tekstituvastuse mootoreid, mida kasutada: Tesseract, OneNote, Google Cloud Vision, OpenCV.

Lahenduse protsess näeb välja selline: imporditakse pilt, pilti töödeldakse, optiline tekstituvastus, sõnade võrdlemine, ettevõtte võrdlemine (otsimine üldregistritest), inimese parandused, salvestamine, saadakse digitaalne arve.

Tulemustena saadi reaalselt palju väiksem ajakulu arve sisestusel (inimene 4,5 minutit ja tarkvara 35 sekundit). Loodud tarkvara andmete õigsuse protsentidena on välja toodud 52%, mis oli vägagi palju allapoole loodetud tulemust (95%). Samas on öeldud, et mitte-leitud väljad on tuletatavad eelneva info pealt ja ainult arve-unikaalsete väljade õigsuse võrdlemises oleks loodetav tulemus 95% saavutatav. Lisaks leiti, et arvete erinevad ülesehitused võivad tekitada probleeme.

Antud töö on väga sarnane plaanitavale tööle. Töös leidus palju asjalikku informatsiooni, kahjuks samuti ka ebatäpset infot. Nimelt on protsessi ja juurutamise pool veidi segane, arve digitaalne formaat ei ole kindlalt paika pandud ning ei tulnud täpselt välja, kas ja kui lihtne oleks võimalik seda siis kasutusele võtta. Samuti jääb andmete õigsus ebaselgeks – miks liialt vähe välju avastati, kas tõlgendati liialt vähe infot või tuvastus ei olnud piisavalt täpne. Taaskord tuleb ka välja, et arvete eri struktuurid tekitavad probleeme ning selles töös erinevaid arveid üldse arvesse ei võetudki.

3.3.4 Lõputöö: ostutšekkide skaneerimise mobiilirakendus

Eestis on 2015. aastal tehtud lõputöö teemal „Mobiilirakendus kulude jälgimiseks ostutšekkide skaneerimise funktsionaalsusega". Selles töös on teostatud sarnane

lahendus, mis tegeleb samuti ostutšekkide töötlusega. Tekstituvastust teostatakse Tesseract'i abil ning tekstist andmete väljalugemisel kasutatakse hajusate regulaaravaldiste teeki FREJ (*Fuzzy Regular Expressions for Java*). Selle eesmärk on kuvada rakenduses kasutajale tema kulusid (ainult kaupmehe nimi ja tšeki summa). [35]

Eelmainitud lõputöö lahendus on väga asjalik ning sarnasel teemal. Töös keskendutakse rohkem aga kogu rakenduse tegemisele ja disainile. Optilise tekstituvastuse osa on teoreetiliselt hästi põhjendatud. Andmete väljalugemise osa on küllaltki väike, sest seal loeti tšekilt välja ainult kaks välja ja see oli teostatud lihtsa muster-reegli abil. Töös on pealiskaudselt ka selle tulemust testitud.

Eelmainitud töö puudusteks on liiga vähese info välja lugemine ja isegi selle puhul madal tulemus. Tehtavas töös on vajalik e-arve koostamiseks palju rohkem infot kuludokumendist välja lugeda ning tõestada, kui hästi on võimalik seda teha. Samuti saadi kindlust, et optilisele tekstituvastusele peab veelgi rohkem rõhku panema kui eelmainitud töös. Seega mingi osa on käesoleva teemaga sama, kuid eesmärk ja väljund on töödel erinevad ning valideerimise osatähtsus on käesolevas töös palju suurem ja täpsem.

4 Prototüüp

Prototüübi sisuks on pildi põhjal automaatne e-arve genereerimine vajalike andmetega. Prototüüpi arendati programmeerimiskeeles Java. Loodi konsoolirakendus, mille sisendiks anti failisüsteemis algne pildistatud pilt ning programm tagastas genereeritud e-arve. Prototüübi lähtekoodi asukoha link on toodud lisas 2.

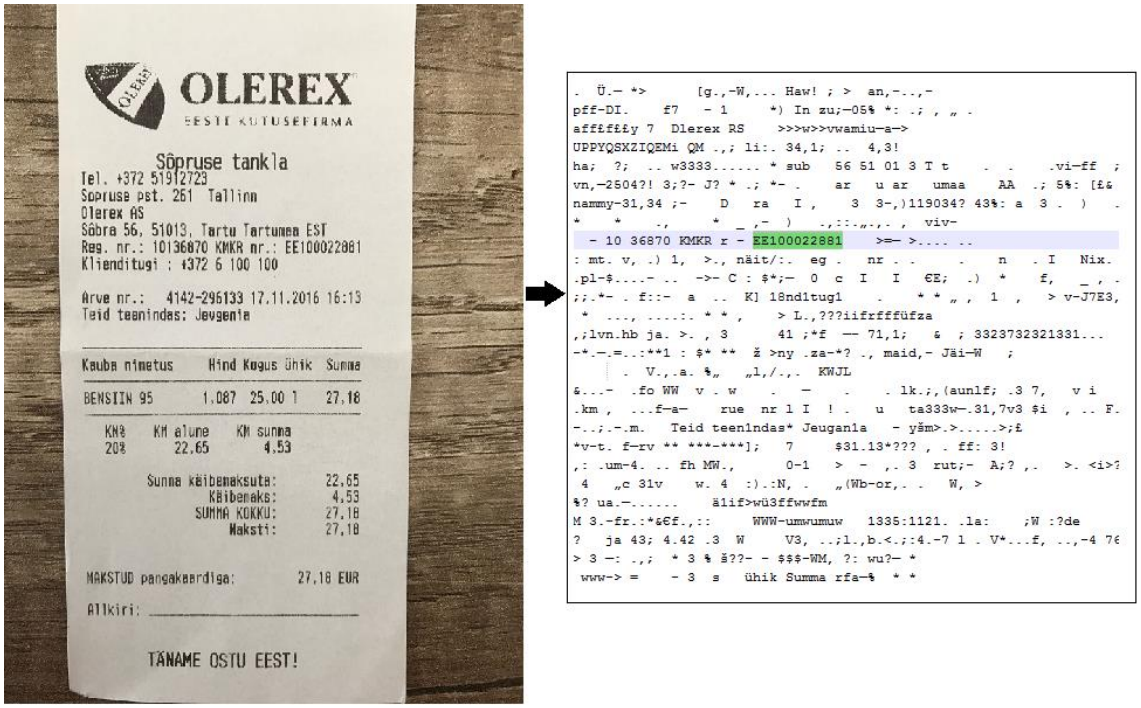
Prototüüpi ei hakatud välja töötama kogu protsessile, mille alla kuuluvad mobiilirakendused, veebiportaal, andmebaas, meiliserverid jpm. Põhjuseks on selliste süsteemide olemasolu. Selle asemel keskenduti töö eesmärgile ja tehti prototüüp just pabertšeki digitaliseerimise osale.

Prototüübi loomisel võeti arvesse teadustöödes soovitatud tehnoloogiaid ja viise. Täpsemad teguviisid, seaded ja loogika leiti autori poolt. Sobilike väärtuste leidmine (näiteks pildi töötlemiste meetoditele etteantavad väärtused) seisnes katsetamisel testvalimi peal ja võrreldes tulemusi. Testvalim koosnes 20’st erineva struktuuri, tausta ja vanusega tšekist (kus oli ka kulunud vanemaid tšekke).

4.1 Optiline tekstituvastus

Optiliseks tekstituvastuseks kasutati käesolevas töös *Tesseract Open Source OCR Engine*'t ja lisati vajalikud keelefailid. Tesseract'ile võtab sisendiks pildifaili ja tagastab selle põhjal tekstifaili. [35]

Optilise tekstituvastuste lahendusi uurides selgus, et töötlemata piltidega ei saa Tesseract hästi hakkama. Selles sai veendutud, kui anti sisendiks mobiilikaamera abil tehtud tšeki pilt ning Tesseracti tulemus on toodud joonisel 4.



Joonis 4. Tesseracti tulemus ilma pilti töötlemata.

Tesseract teeb enne teksti tuvastust küll mõningat pilditöötlust, aga sellest silmnähtavalt ei piisa. Selleks otsustati pilti eelnevalt Tesseract'ile sobivamaks töödelda.

4.1.1 Pildi töötlemine enne OCRi

Pildi töötlemiseks kasutati avatud lähtekoodiga teeki OpenCV, mille sisuks on arvutinägemine ja masinõpe [36] [37].

4.1.1.1 Pildi pööramine

Eeldatakse, et sisendpilt on tehtud püstises asendis, sest tavaliselt on tšekid kujult piklikud. Osadel telefoniga tehtud piltidel ei ole automaatne asetus õige, mis võis tuleneda pildistamise asendit (ülevalt alla ja seetõttu ei määratud õiget asendit).

Prototüübis kontrollitakse pildi asetust ning pööratakse pilt vajadusel püstisesse asendisse, milleks kasutatakse OpenCV *rotate* meetodit.

4.1.1.2 Automaatne kärpimine

Tšekkide pikliku kuju tõttu jääb pildistades taust näha. Kui taust on värviline või kirju, siis kulub pildi töötluseks väga kaua aega ja Tesseract OCR tuvastab äärtest üleliigseid sümboleid. Seetõttu eemaldatakse pildilt taust nii, et jääks alles ainult tšeki ala.

Kärpimise jaoks tehakse järgnevad sammud:

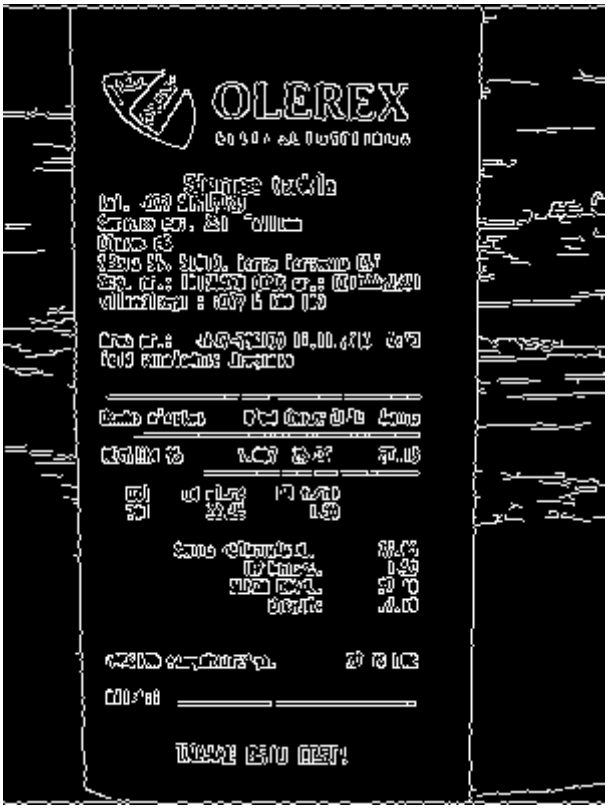
1. Suuruse kahandamine
    a. Tegevus: pildi kõrgus kahandatakse 400 pikslini ja laius proportsionaalselt.
    b. Eesmärk: väiksem pilt ei ole nii terav ja kontuure on kergem leida.
2. Alla ja üles joonte lisamine
    a. Tegevus: ülesse ja alla äärde lisatakse 2 piksli suurused värvilised jooned
    b. Eesmärk: kui tšekki on pildistatud nii, et maksimaalne võimalik osa jääb peale (taust ainult vasakul ja paremal), siis ei leita pildilt sisemisi kontuure. Seega lisati valgest (tavaliselt tšeki värvist) erinevat värvi jooned.

3. Äärte märgistamine

    a. Tegevus: märgitakse pildile ääred, kasutades *Canny* äärte detektorit

    b. Eesmärk: saamaks eristatava pinna ehk tšeki ääred

    c. Tulemus on toodud joonisel 5.

i. 

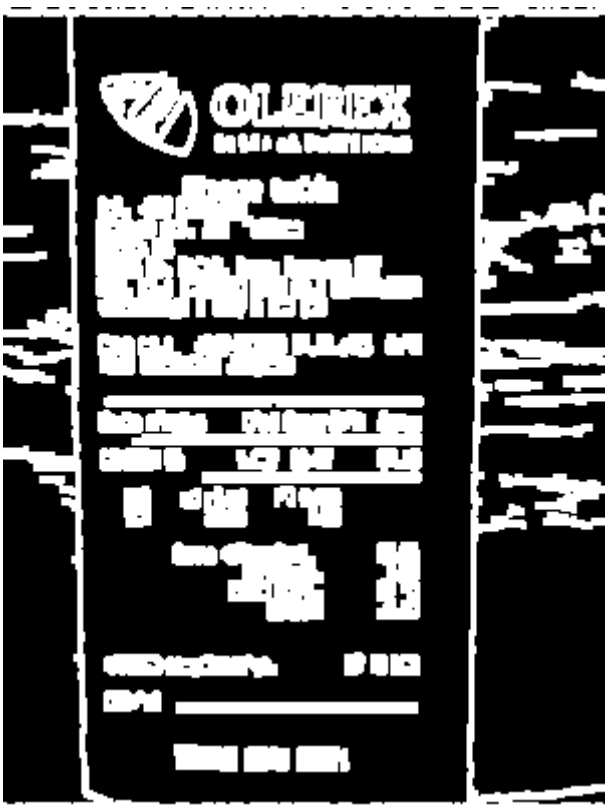
Joonis 5. Pildi töötlemine: äärte märgistamine.

4. Äärte laiendamine

    a. Tegevus: laiendatakse leitud ääred kasutades *dilate* meetodit.

    b. Eesmärk: kirju tausta puhul võivad ääred olla katkenud, neid laiendades saab selgelt eristatava tšeki ala (ristküliku)

    c. Tulemus on toodud joonisel 6.

i. 

Joonis 6. Pildi töötlemine: äärte laiendamine.

5. Kontuuride leidmine

    a. Tegevus: *findContours* abil leitakse pildilt kontuurid, millega saadakse kõikide kontuuride alad.

    b. Eesmärk: saamaks kätte eristatavad objektid

    c. Tulemus on toodud joonisel 7.

i. 

Joonis 7. Pildi töötlemine: kontuuride leidmine.

6. Tšeki ala leidmine kontuuride seast

    a. Tegevus: leitud alad pannakse listis suuruse järjekorda, seejärel kontrollitakse suurima ala suurust (peab olema kogu pildi alast vähemalt 20%, mis on minimaalne ala väga pikkade tšekkide puhul)

    b. Eesmärk: saamaks tšeki konkreetse ala punktid

7. Esialgse pildi kärpimine

    a. Tegevus: saadud tšeki ala punktidest moodustatakse ristkülik kasutades *boundingRect* meetodit. Ristküliku element teisendatakse tagasi vastavusse esialgse pildi mõõtudele. Ristküliku ala järgi kärbitakse esialgset pilti.

    b. Eesmärk: saamaks esialgse õiges suuruses pildist tausta eemaldatud

    c. Tulemus on toodud joonisel 8.

Joonis 8. Pilt peale automaatset tausta eemaldamist.

Kui tšeki ala ei leitud (näiteks valge tausta tõttu), siis jätkatakse töötlemist esialge mitte-kärbitud pildiga.

4.1.1.3 Konvertimine must-valgeks

Pilt konverditi must-valgeks OpenCV abil. See võimaldab pildi kiiremat töötlust ja on vajalik järgnevate pilditöötlusprotsesside jaoks. Joonisel 9 on näha pilt must-valgena.

Joonis 9. Pildi töötlemine: must-valge pilt.

4.1.1.4 Pildi binaarne kuju

Viimasena tehakse pildile lävisegmentimine, mis aitab must-valgest pildist tekitada binaarse pildi. Binaarsel pildil on iga piksli jaoks ainult kaks võimalikku väärtust, seega muudetakse selle abil tähemärkide ääred konkreetsemaks.

Prototüübis on kasutatud kahte erinevat viisi:

5. Kärbitud pildi puhul (pildi ala ning musta ja valge suhte erinevus väiksem):

    a. Adaptiivne lävisegmentimine

        Kasutades meetodit *adaptiveThreshold* ja Gaussi jaotust. Adaptiivne lävisegmentimine võtab arvesse ka pildi valguse erinevusi (aladel erinevad varjud) ja läved arvutatakse pildist jaotatud väiksematele aladele eraldi.

6. Kärpimata pildi puhul (pildil hele taust, valge osakaal mustast suurem):

    a. Tavaline lävisegmentimine

        Kasutades meetodit *threshold* ja Otsu binariseerimist, mis automaatselt arvutab lävisegmentimise väärtuse vastavalt etteantud pildile.

Joonisel 10 on toodud lävisegmentimise tulemus.


OLEREX
EESTI KÜTUSEFIRMA

Sõpruse tankla
Tel. +372 51912723
Sopruse pst. 261 Tallinn
Olerex AS
Sõbra 56, 51013, Tartu Tartumaa EST
Reg. nr.: 10136870 KMKR nr.: EE100022881
Klienditugi : +372 6 100 100

Arve nr.: 4142-296133 17.11.2016 16:13
Teid teenindas: Jevgenia

Kauba nimetus	Hind	Kogus	ühik	Summa
BENSIIN 95	1,087	25,00	l	27,18

KM%	KM alune	KM summa
20%	22,65	4,53

Summa käibemaksuta:	22,65
Käibemaks:	4,53
SUMMA KOKKU:	27,18
Maksti:	27,18

MAKSTUD pangakaardiga: 27,18 EUR

Allkiri: _____

TÄNAME OSTU EEST!

Joonis 10. Pildi töötlemine: lävisegmentimise tulemus.

4.1.2 Tesseract'i tekstituvastus

Tesserct'i tekstituvastuse sisendiks on pilt ja väljundiks pildi põhjal tuvastatud tekst. Samuti on võimalik optilise tekstituvastuse mootorile ette anda erinevaid konfiguratsioone.

Prototüübis kasutatakse järgnevaid seadeid:

- Keeled: eesti + inglise

  Ainult eesti keel ei andnud vastu valimit piisavalt jäid tulemust. Inglise keele tugi on Tesseract'il kõige parem ning eesti keelele lisaks seati ka inglise keel lubatud keele parameetrina.

- Lubatud tähemärgid on toodud joonisel 11.

```
private static final String TESS_ALLOWED_CHARS =
"ABCDEFGHIJKLMNOPQRSTUVÕÜÄÖWXYZŽŠabcdefghijklmnopqrstuvõüäöwxyzžš01234
56789.,%-–():/\\?!;@£$€[]& *<>=";
```

Joonis 11. Tesseract'ile tuvastamiseks lubatud tähemärgid.

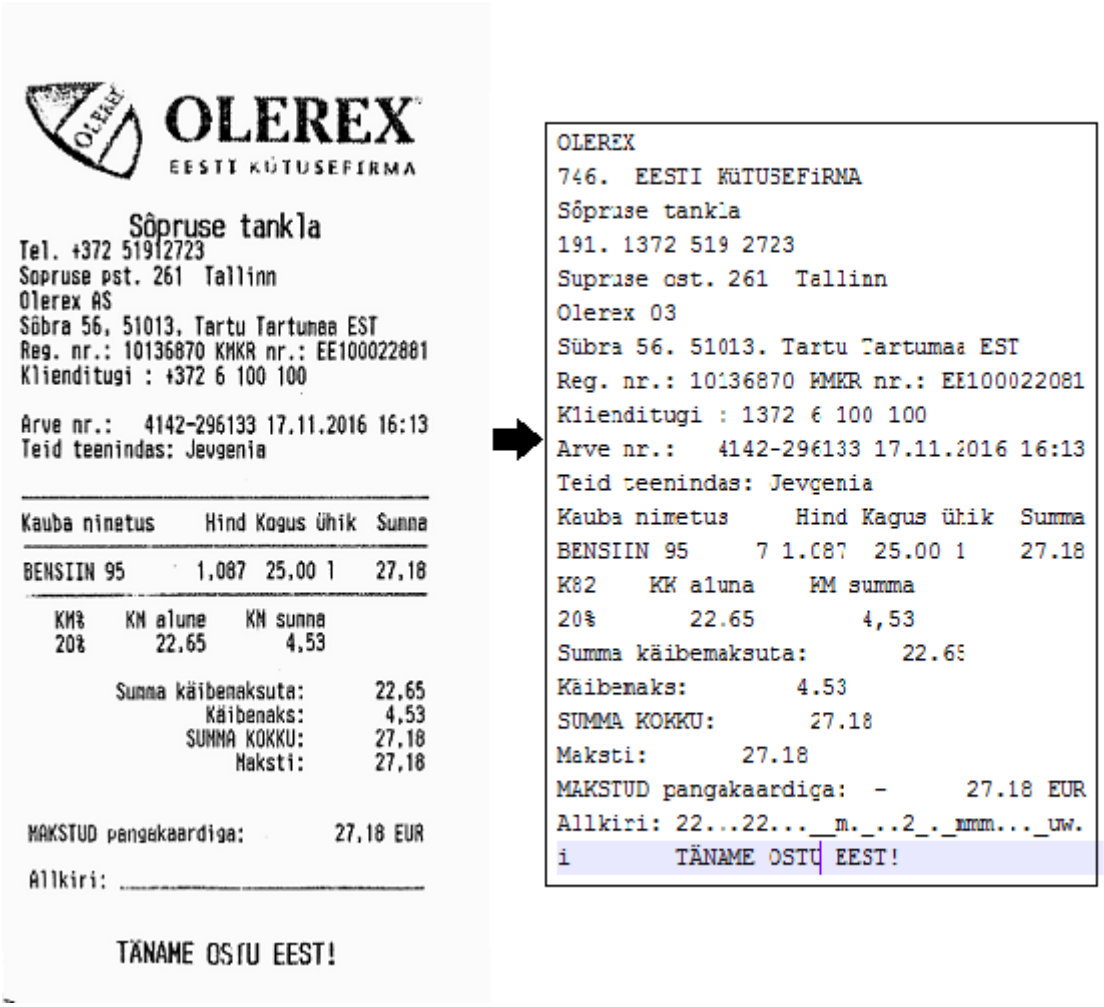
  Selle abil tuvastatakse piltidelt ainult ülal mainitud tähemärke, milleks on eelkõige väiksed ja suured tähed ning numbrid, lisaks ka summade jaoks kasutusel olevad sümbolid. Lubatud sümbolite kasutamine aitas tulemustest eemaldada mitte-standardsed sümbolid ja saada adekvaatsemaid tulemusi.

- Teksti struktuuri säilitamine

  Vaikimisi Tesseract pildi struktuuri ei säilita ehk tekst loetakse rea haaval sisse ning sõnade vahel on säilitatud ainult üks tühik. Tesseract'i parameeter *preserve_interword_spaces* lubas säilitada teksti struktuuri ehk tulemuses on sõnad eraldatud mitmete tühikutega ja tekstikujul tulemus sarnaneb rohkem tšeki omale. See aitas paremini infot eristada, sest tšekkidel on ühel real tihti kuvatud erinevaid andmeid.

4.1.2.1 Tekstituvastuse tulemus

Peale pildi eeltöötlust ja Tesseract'i seadete seadmist saadi tekstituvastuse tulemusena palju parem tulemus kui alguses, mis on kujutatud joonisel 12.



Joonis 12. Tesseract'i optiline tekstituvastus töödeldud pildiga.

Joonisel 12 on näha vasakul töötlemise tulemus ja paremalt Tesseract OCR poolt tuvastatud tekst.

4.2 Väljade märgistamine

4.2.1 Tšekkide struktuur

Kõige laialdasemalt levinud tšeki struktuur ülevalt alla suunas on järgnev:

1. Kaupmehe andmed:
    a. Registrikood
    b. Nimi
    c. KMKR
    d. kontaktandmed
2. Tšeki andmed:
    a. Number
    b. Kuupäev
3. Tšeki sisu (toodete või teenuste list):
    a. Nimi
    b. Kogus
    c. Hind
4. Kokkuvõttev summade sektsioon:
    a. Toodete/teenuste hinnad summeeritult
    b. Maksude info (maksu protsent, netosumma, maksu summa, lõplik summa)
5. Tehingu info vastavalt maksevahendile (pangakaardi numbrid jms)

Tšeki pildil, mis on kujutatud joonisel 13, on värviliselt tähistatud kõige tähtsamad andmed ehk vajalikud väljad, mida raamatupidamise kuludokumendi sisestuseks vaja läheb.

Joonis 13. Tšeki pilt koos visuaalse andmeväljade kujutamisega.

4.2.2 Reeglipõhine

Tekstist andmete leidmiseks kasutati reeglipõhist lahendust, sest tegemist on kitsa ja konkreetse valdkonnaga. Tšekkide struktuur veidi erineb, kuid vajalik andmehulk on igal tšekil olemas ning neid on võimalik reeglite abil üles leida.

Prototüübis realiseeriti reeglipõhine lahendus kasutades regulaaravaldisi. Selle jaoks loodi järgnev struktuur:

- Reeglite kogum

  Reeglite kogum sisaldab andmeobjektide regulaaravaldisi. Kasutusel on lihtsad kui ka grupeerimistingimustega avaldised, ühe- kui ka mitmerealised avaldised ning ka avaldiste listid, milles on elemendid tähtsuse järjekorras.

- Vastete leidmine

  Sisaldab üldist funktsionaalsust leidmaks tekstist regulaaravaldise vasted. Sisendina saab anda ühe konkreetse regulaaravaldise või –avaldiste listi ning samuti keerulisema regulaaravaldise puhul grupi indeksi (mis on kasutusel saamaks regulaaravaldise tulemusest konkreetset defineeritud osa). Väljundina tagastatakse esimene vaste või kogu vastete list.

  Sisendina antud regulaaravaldiste list võimaldab defineerida reeglite tähtsuse. List käiakse järjekorras läbi ning tagastatakse esimene tulemus või tulemuste list vastavalt reeglite järjekorrale.

- Valideerimine

  Selle funktsionaalsuse käigus otsustatakse lõplik tulemus. Esmalt leitakse reeglite järgi vasted ning seejärel valideeritakse saadud info. Valideerimisena on näiteks kasutusel kuupäevade õigsuse, koodide kontrollsumma, summade seose kontrollid.

Järgnevalt on toodud täpsemad kirjeldused, mis reeglite abil leiti tekstist üles erinevad andmeväljad.

4.2.2.1 Kaupmehe andmed

4.2.2.1.1 Registrikood

Igal Eesti ettevõttel on registrikood, mis on 8 tähemärgi pikkune ja koosneb numbritest. Algselt leiti tekstist 8-kohalisi numbrilisi koode, mille reegel on toodud joonisel 14.

```
REGNO_FORMAT = "\\b[0-9]{8}\\b";
```

Joonis 14. Koodinäidis: registrikoodi reegel.

Ainuüksi sellise reegli tulemustele ei saanud kindel olla, sest erinevaid 8-kohalisi numbreid võib tšekil olla mitmeid. Seetõttu kontrollitakse leitud 8-kohalise numbri valiidsust. Eesti ettevõtete registrikoodi on võimalik valideerida, sest see kood sisaldab kontrollnumbrit. Seega kontrollitakse leitud tulemuse õigsust ja alles seejärel omistatakse tšeki objektile müüja registrikood.

4.2.2.1.2 Käibemaksukohustuslase number

Kui ettevõte on käibemaksukohustuslane, siis on tal käibemaksukohustuslase identifitseerimisnumber. Sel numbril on kindel formaat: number algab kahekohalise maakoodiga, millele järgneb 2-12 numbrit või tähte.

Eesti käibemaksukohustuslase numbri (KMKR) formaat on “EE” + 9 numbrit [38].

Prototüübis otsitakse pildi põhjal tuvastatud tekstist regulaaravaldise teel Eesti KMKR’i formaati, mille reegel on toodud joonisel 15.

```
VAT_FORMAT = "\\b[E]{2}[0-9]{9}\\b";
```

Joonis 15. Koodinäidis: käibemaksukohustuslase numbri reegel.

4.2.2.1.3 Ettevõtte nimi

Ettevõtte nime otsitakse tekstist ettevõtte tüübi lühendi järgi. Joonisel 16 on kujutatud selle reegel.

```
SELLER_NAME_FORMAT = "[A-Z](.*) (\\b(?i)(AS|OÜ|UÜ|MTÜ|TÜ|FIE)\\b)";
```

Joonis 16. Koodinäidis: ettevõtte nime reegel.

4.2.2.2 Tšeki sisu

4.2.2.2.1 Arve/kviitungi number

Ostutšeki sisestuseks on vaja tšekile viitamiseks identifikaatorit. Üldiselt on tšekil välja toodud arve number ja ka kviitungi number. Arve number on tavaliselt seotud tehinguga

ning kviitungi number kaardimaksega. Mõnel ostutšekil arve number puudub ja on mainitud see kviitungi nime all.

Tšeki numbri leidmisel kasutati listi, et esimesena leida arve number. Selle puudumisel leitakse kviitungi number. Joonisel 17 on toodud selle leidmise reegel.

```
INVOICE_NO_FORMAT = new String[]
            {"(?m)^(.*\\b(?i)(Arve)\\D*)(\\d+\\S*)\\s",
               "(?m)^(.*\\b(?i)(Kviitung)\\D*)(\\d+\\S*)\\s"};
```

Joonis 17. Koodinäidis: tšeki numbri reegel.

Otsingut alustatakse algusest, sest tavaliselt on ostutšeki ülal müüja info ning tagastatakse esimene vaste.

4.2.2.2.2 Kuupäev

Kuupäev leitakse tekstist kuupäeva formaadi otsinguga. Eelkõige otsitakse tavapärast kuupäeva formaati, mille eraldajaks punkt. Regulaaravaldisse on lisatud eraldajaks punkti juurde ka koma, sest optiline tekstituvastus võib koma ja punkti kergelt segamini ajada. Järgnevana proovitakse leida ka selliseid kuupäevi, kuid eraldajaks sidekriips või kaldkriips. Joonisel 18 on toodud need reeglid koodi kujul.

```
DATE_FORMAT = new String[]
{"\\b\\d(\\d)?(\\d\\d)?(\\.|,)\\d(\\d)?(\\.|,)\\d\\d(\\d\\d)?\\b",
 "\\b\\d(\\d)?(-|/)\\d(\\d)?(-|/)\\d\\d(\\d\\d)?\\b",
 "\\b\\d\\d\\d\\d(-|/)\\d(\\d)?(-|/)\\d\\d\\b"};
```

Joonis 18. Koodinäidis: kuupäeva reegel.

Kuupäeva õigsus valideeritakse teksti konverteerimisel Java kuupäeva tüübiks, millele on etteantud kuupäeva formaadid, mis on toodud joonisel 19.

```
DATE_PATTERNS = new String[] {
        "dd.MM.yyyy",
        "dd.MM.yy",
        "MM.dd.yy",
        "yy.MM.dd",
        "yy.dd.MM",
        "yyyy.MM.dd",
        "yyyy.dd.MM"};
```

Joonis 19. Koodinäidis: kuupäeva valideerimise formaadid.

4.2.2.2.3 Summad

Summad on ostutšekil kirjas alumises osas, seega otsitakse summasid ainult optilisest tekstituvastusest saadud teksti alumisest osast. Selle arvelt hoitakse kokku regulaaravaldiste leidmise aega ja mitte-valiidsete vastete töötlemist.

4.2.2.2.3.1 Kogusumma

Ostutšeki kogusumma leidmiseks otsitakse summade sektsiooni enim levinud nimetuste järgi. Selle leidmisel võetakse töötlusse käesoleva ja ka järgmine rida. Joonisel 20 on kujutatud reeglit programmikoodina.

```
SUM_FORMAT =
"(?m)^.*(?i)(kokku|summa|total|maksti|makstud|pangakaart|tasuda|maksta).*\\n.
*$";
```

Joonis 20. Koodinäidis: kogusumma reegel.

Nendelt kahelt realt otsitakse lõppsummat ettemääratud summa formaadi regulaaravaldise teel. Sealjuures välistatakse vana Eesti Krooni valuutaga olevad summad, sest osadel tšekkidel on informatiivse osana summa kuvamine Eesti Kroonides siiani kasutusel. See reegel on kujutatud joonisel 21.

```
AMOUNT_FORMAT =
        "\\b(\\d+(\\.|,)(\\d+)?)(\\D|$)(?!" + OLD_CURRENCY + ")";
```

Joonis 21. Koodinäidis: summa üldise formaadi reegel.

4.2.2.2.3.2 Netosumma

Netosumma leidmisel otsitakse algul samuti nimetuse järgi read välja, mille reegel on joonisel 22..

```
NET_SUM_FORMAT =
"(?m)^.*(?i)(netosumma|ilma|kokku|summa|total|maksti|makstud).*\\n.*$";
```

Joonis 22. Koodinäidis: netosumma reegel.

Seejärel otsitakse samuti ülalmainitud summa formaadi järgi vastet.

4.2.2.2.3.3 Käibemaksu määr

Käibemaksu protsendi leidmisel otsitakse kuni kahekohalist arvu, mis võib olla tuhandiku täpsusega ja lõppema protsendi-märgiga. Reegel on toodud joonisel 23.

```
VAT_AMOUNT_FORMAT = new String[]
                   {"\\b([1-2]{1}[0-9]{1})\\b\\s?(%)",
                    "\\b([1-2]{1}[0-9]{1})(\\.|,)(0)?(0)?\\b\\s?(%)"};
```

Joonis 23. Koodinäidis: käibemaksu määra reegel.

4.2.2.2.3.4 Käibemaksu summa

Käibemaksu summa leidmisel otsitakse algul samuti nimetuse järgi ridu, mille reegel on toodud joonisel 24.

```
VAT_TEXT_FORMAT = "(?m)^.*(?i)(KM|käibemaks|VAT|).*\\n.*$";
```

Joonis 24. Koodinäidis: käibemaksu summa reegel.

Seejärel otsitakse samuti ülalmainitud summa formaadi järgi vastet.

4.2.2.2.3.5 Valuuta

Valuuta leitakse alumise osa tekstist kolmetähelise suurtes tähtedes sõna järgi, mille reegel on joonisel 25.

```
CURRENCY_FORMAT = "\\b[A-Z]{3}\\b";
```

Joonis 25. Koodinäidis: valuuta reegel.

Leitud tulemust kontrollitakse, kas leitud tulemus on kasutusel olevate valuutade hulgas. Selleks kasutatakse Java *util*'is olevat klassi *Currency*.

4.2.2.2.3.6 Summade valideerimine

Leitud summade puhul kontrollitakse nende õigsust arvutuste teel.

Summade seosed on järgnevad:

7. Kogusumma = netosumma + käibemaksu summa

8. Käibemaksu summa = netosumma * käibemaksu määr (protsent)

Kui summad on valiidsed, siis määratakse tšeki netosumma, käibemaksu summa, käibemaksu protsent ja kogusumma.

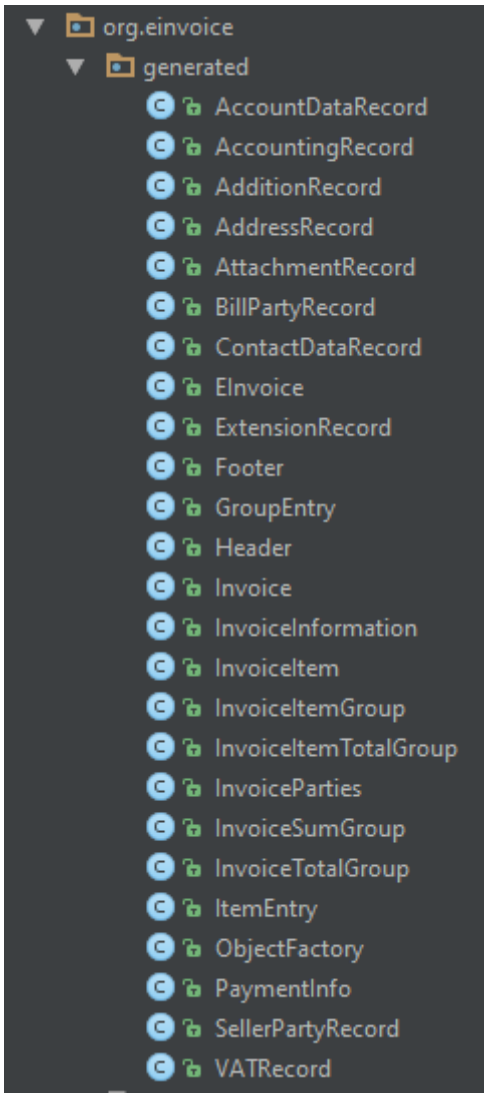
4.3 E-arve genereerimine

Prototüübis on realiseeritud e-arve genereerimise funktsionaalsus. Vastavalt defineeritud sisendandmetele väljastab prototüüp XML formaadis e-arve, mis on kooskõlas e-arve nõuetega.

4.3.1.1 E-arve standardformaat Java koodiks

E-arve on XML vormingus dokument, millel on eksisteerib skeemi kirjeldav XSD dokument. [18]

Prototüübis kasutatakse e-arve versiooni 1.2. Vastavalt e-arve formaadi kirjelduse dokumendile (XSD) genereeriti JAXB abil Java klassid, mis on nähtaval joonisel 26 [39].



Joonis 26. Genereeritud E-arve Java klassid.

4.3.1.2 E-arve genereerimine andmete põhjal

Genereeritud e-arve klasse kasutades saab defineerida e-arves oleva info. Sisendiks kasutatakse ostutšeki digitaliseerimise tulemusena tšeki andmetega objekti, millel on järgnevad andmed:

1. Ostja info
    a. Registrikood

    b. Nimi
    c. Käibemaksukohustuslase number
2. Müüja info
    a. Registrikood
    b. Nimi
    c. Käibemaksukohustuslase number
3. Tšeki number
4. Tšeki kuupäev
5. Summad
    a. Kogusumma
    b. Netosumma
    c. Käibemaksu määr
    d. Käibemaksu summa
    e. Valuuta
6. Ostutšeki pilt

Prototüüp võimaldab nende andmetega genereerida valiidse e-arve XML faili.

4.3.1.2.1 E-arvele manuse lisamine

Vastavalt raamatupidamise seaduse §12 lõige 1 on raamatupidamiskohuslane kohustatud säilitama raamatupidamise algdokumenti 7 aastat. Sama seaduse §9 lõige 9 sätestab, et algdokumenti lubatud üle viia teise vormingusse, kui üleviimise käigus ei muudeta majandustehingut puudutavaid andmeid [16]. Hetkel teevad raamatupidajad igast ostutšekist manuaalselt koopia, sest ostutšekkide termotrükk ei säili 7 aastat. Eelmainitu annab kinnitust sellele, et e-arvele algdokumendi juurde lisamine on lubatud.

E-arve toetab manuse lisamist, mis peab olema PDF-formaadis ja BASE64 kodeeringus. Ostutšeki koopia on meil pildifaili kujul olemas, mis oli aluseks tšeki digitaliseerimiseks.

Pildifaili konvertimiseks PDF kujule kasutati Java teeki *itext*, mis võimaldus luua PDF faili ning selle sisuks seada pildi [40]. Seejärel kodeeriti pilt kasutades BASE64 kodeerimismeetodit ning lisati saadud tulemus e-arvele.

5 Valideerimine

Käesoleva töö raames valmis ostutšekkide digitaliseerimise prototüüp. Selle abil valideeriti ka töös seatud eesmärke. Valideerimisel kasutati võrdluseks manuaalset kuludokumentide sisestamise protsessi infot, mis saadi raamatupidajatelt.

5.1 Testimine

Testimiseks sooviti katsetada prototüüpi võimalikult erinevate olukordade puhul. Tulemuste valideerimiseks väljastati võimalikult palju infot digitaliseerimise protsessi kohta.

Pilditöötlus ja optiline tekstituvastus on üsnagi ressursikulukad, seega mainitakse töös ära keskkonna parameetrid. Prototüüp töötas töö autori sülearvutis, mille tehnilised andmed on järgnevad: protsessor Intel I5-3317U 1.70GHz, muutmälu 8GB, tavaline kõvaketas 500GB. Java programmi arendati ja käivitati integreeritud programmeerimiskeskkonna Intellij IDEA abil.

5.1.1 Protsess

Testimiseks valiti välja 90 ostutšeki pilti, mis varieerusid suuruse (ruudud ja piklikud ristkülikud) ja ülesehituse (väljade asetus ja info hulk) poolest. Tšekke oli erinevate kaupmeeste omasid (nii kütusefirmade, ehitus-, tehnika kui ka söögipoodide omasid). Pea kõik tšekid olid voltimisjoontega. Samuti oli nende hulgas kortsus ja ka mõni vanem tšekk (1-2 aasta vanune), mille tekst oli juba veidi tuhmunud.

Välja valitud tšekke pildistati mobiiltelefoniga iPhone 7 (kaamera 12 megapikslit) ja iPhone 5S (kaamera 8 megapikslit) abil. Pildid tehti erineva valguse käes (päevavalgus ja lambivalgus) ning erinevate taustade peal (pooled valge taustaga ning ülejäänud kirju pruunika taustaga).

Prototüüpi lisati silmus, mis käiks automaatselt järjest eelmainitud 90 pildifaili läbi. Igal pildifaili puhul käivitati digitaliseerimise protsess. Hilisema valideerimise jaoks oli vaja protsessi erinevate osade kohta saada rohkem infot.

Digitaliseerimise käigus salvestati maha järgmised failid iga tšeki kohta:

- Töödeldud pilt

  Esialgse pildiga võrreldes on võimalik veenduda, kas pilt kärbiti korrektselt ja kas tekst on muudetud loetavaks.

- OCR poolt tuvastatud tekst

  Esialgse pildiga võrreldes on võimalik veenduda, kas tuvastatud tekst on korrektne.

- Java objekti *Receipt* ehk tšeki väärtuste väljatrükk

  Tuvastatud tekstiga võrreldes on võimalik veenduda, mis määral reeglid õiged tulemused.

  Väljatrüki näidis on toodud joonisel 27.

```
===RECEIPT===
Seller's reg no: 11199045
Seller's KMKR:   EE101018359
Seller's name:   Klick Eesti AS
Receipt ID:      0000000232000022127
Receipt date:    17.03.17
(net amount:     74.99)
(vat amount:     15.00)
VAT rate:        20
Total amount:    89.99
Currency:        EUR
```

Joonis 27. Tšeki väärtuste väljatrüki näidis.

- E-arve XML fail

  Võimalik valideerida, kas väljatrükis toodud andmed on korrektselt e-arves väärtustatud. Lisa 3 all on toodud näidis ühest prototüübi poolt väljastatud korrektsest e-arvest.

5.1.2 Automaatne statistika

Prototüüpi loodi lisaks automaatne statistika koguja, et vähendada manuaalse kontrollimise ja võrdlemise tööd.

Selle abil mõõdetakse ühe tšeki digitaliseerimisele kulunud aega, mis aitab ka identifitseerida, milline pilt oli teistest ajakulukam. Peale kõikide failide töötlemist arvutatakse välja keskmine protsessimise aeg.

Järgmisena kogutakse statistikat iga tšeki leitud andmete kohta. Selleks kontrollitakse Java klassi *Receipt* ja talle omistatud väärtusi. Selle tulemusena saadakse statistika, mitmel korral iga välja kohta digitaliseerimise protsessi käigus tulemus leiti.

Lõpuks väljastatakse statistika tekstifailina (statistika näidis toodud lisas 4).

5.1.3 Tulemused

5.1.3.1 Andmete leidmine

Tabelis 1 on toodud tšeki objektile andmeväljade väärtustamise tulemused.

Tabel 1. Tšeki andmeväljade väärtustamise tulemused.

Andmeväli	Leitud väärtuseid (tk)	Leitud väärtuseid (%)
Registrikood / KMKR	83 / 90	92 %
Müüja nimi	42 / 90	47 %
Kuupäev	68 / 90	76 %
Tšeki number	67 / 90	74 %
Netosumma	69 / 90	77 %
Käibemaksu summa	40 / 90	44 %
Käibemaksu määr	70 / 90	78 %
Valuuta	65 / 90	72 %
Kogusumma	81 / 90	90 %

Enim leiti tekstist kaupmehe registrikoodi / KMKR'i. Nende väljade peal on valiidsuse kontroll tõhus ning 92% on ülalmainitud testhulga kohta hea tulemus. Samuti väärtustati 90% juhtudel kogusumma väärtus, mille õigsus kontrolliti manuaalselt üle.

Madala väärtustamise tulemusena on müüja nimi ja käibemaksu summa (vastavalt 47% ja 44%). Kuna käibemaksu määr leiti 78% juhtudel, siis oleks võimalik arvutuste teel tõsta käibemaksu summa väärtustamine samale tasemele. Samuti on võimalik müüja nimi pärida välistest registritest registrikoodi alusel, mis samuti tõstaks selle välja väärtustamist lausa 92 protsendini.

Tšeki number ja kuupäev on pigem unikaalsemad väärtused, mida teistest väljadest tuletada ei ole võimalik. Nende väärtustamise protsendid on vastavalt 74 ja 76. Nendest

madalaimat tulemust ehk 74% võib seega lugeda üleüldiseks tšeki väärtustamise protsendiks, sest ülejäänud vajalike ja tuletatavate andmeväljade leidmise tulemus on kõrgem. Vajalike väljade keskmine leidmise tulemus oli 82%.

5.1.3.2 Andmete õigsus

Lisaks andmeväljade väärtustamise tulemustele kontrolliti ka andmete õigust. Andmete õigsust kontrolliti üle manuaalselt e-arve ja pildi võrdluse teel. Täpsemalt kontrolliti minimaalselt vajalike andmete väärtuste õigsust, milleks olid ettevõtte registrikood/KMKR, kogusumma, tšeki number, kuupäev ja käibemaksu määr. Nende väärtused olid 91% juhtudel õigesti välja loetud. Kõige rohkem oli probleeme tšeki numbri välja lugemisega – osadel juhtudel oli loetud poolik number või oli aetud tähed ja numbrid omavahel segamini). Põhjus seisneb selles, et tšeki numbril ei ole kindlat formaati ja seetõttu ei ole seda võimalik valideerida. Raamatupidajatelt saadud info kohaselt puudub mõnel juhul tšeki peal arve number üldse ning tähtis on saada raamatupidamisprogrammi sisestatud tšeki pealt mõni unikaalne väli, et seda hiljem algdokumendiga siduda.

Väärtusi ei avastatud üldse järgmistel juhtudel:

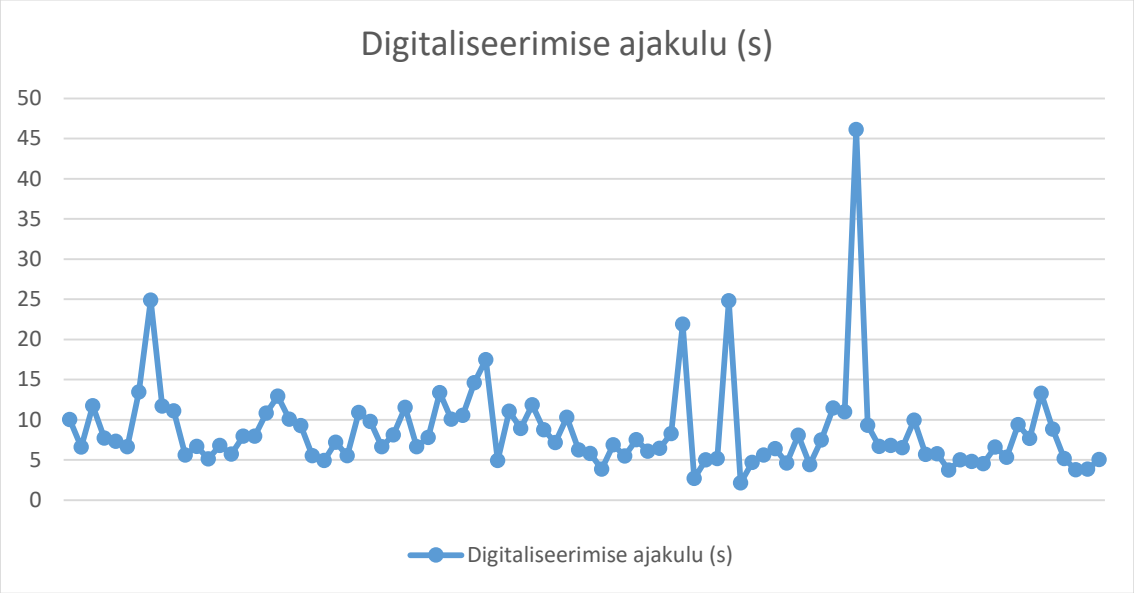
- Tšeki tekst oli liigselt kulunud, mida tingis termotrüki vanus (sest testandmetes oli 1-2 aasta vanuseid tšekke).
- Tšekil olev liiga väike kiri ja kirja suuruse erinevused (neil juhtudel ei tuvastanud OCR teksti piisavalt hästi)
- Kirjavead optilise tekstituvastuse tulemusena. Näiteks sõna “arve” oli väga mitmel korral tuvastatud nagu “nrua”, samuti oli segamini aetud numbrid 5 ja 6.

Samuti kontrolliti veel eraldi 8% juhtumeid, kus müüja registrikoodi või käibemaksukohustuslase numbrit ei leitud. Tulemusena avastati, et probleem oli pigem ekstreemsetes sisendandmetes ehk kulunud ja kortsus tšekkides.

5.1.3.3 Digitaliseerimise ajakulu

Joonisel 28 on toodud prototüübi poolt kulunud aeg iga tšeki digitaliseerimiseks. Keskmiseks digitaliseerimise ajakuluks oli 8 sekundit. Minimaalne ajakulu oli 2

sekundit. Maksimaalne ajakulu oli ühe tšeki puhul 46 sekundit, järgmine ajakulukas digitaliseerimine oli 25-sekundiline.



Joonis 28. Digitaliseerimise ajakulu diagramm.

Joonisel 29 on maksimaalse ajakuluga tšekiks osutunud pilt, millel on kirju taust ning tšekil olev teksti suurus varieeruv. Kirjut tausta ei suutnud prototüüp eraldada, sest ei leitud korrektset ristkülikut (tšekil oli kviitung juurde lisatud, mis voltides tekitas 6-nurkse kujundi.

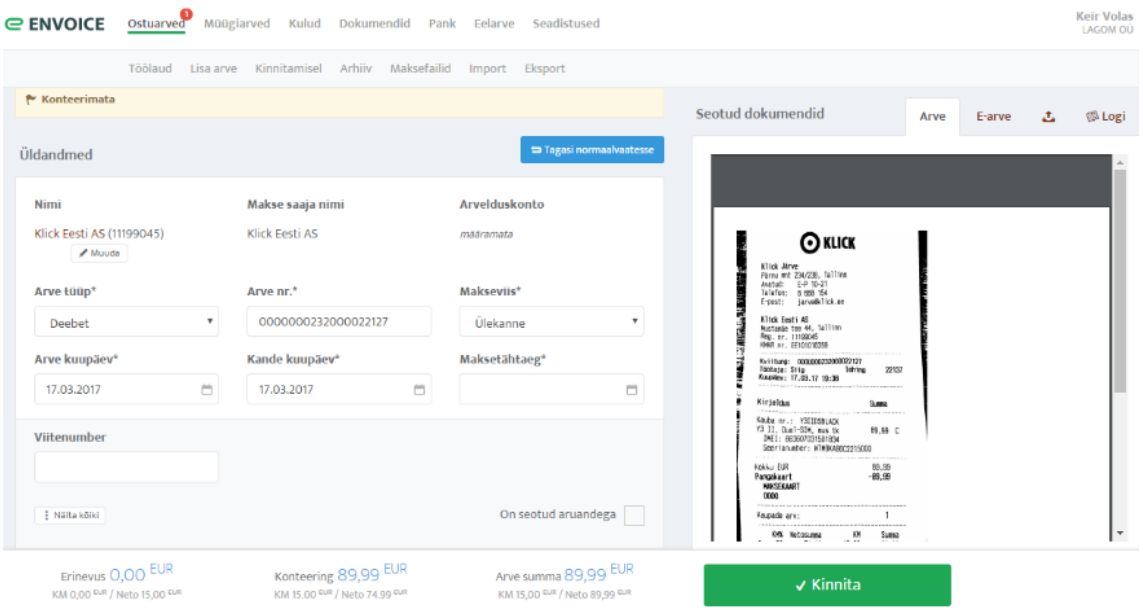
Joonis 29. Maksimaalse töötlemise ajakuluga tšeki pilt.

5.2 E-arvete õigsuse kontroll

E-arve XML formaadil on defineeritud XSD failiga skeemi kirjeldavad reeglid. Tänu XSD failile on võimalik XML valideerida. Valideerimine teavitab probleemidest, näiteks kui struktuur on vigane, kohustuslikud elemendid puudu või tegelikud väärtused ei vasta väljade reeglitele.

Prototüübi poolt korrektselt tuvastatud ostutšeki e-arve valideeriti vastu internetis olevat validaatorit [41]. Tulemus oli positiivne: prototüübi poolt genereeritud e-arve XML fail oli valiidne.

Järgmine õigsuse kontrolli samm oli faili importimine raamatupidamisalasesse tarkvarasse, saamaks kindlust, kas andmed tulid läbi ja kõik info parsitakse korrektselt. Üldiselt liiguvad e-arved raamatupidamisprogrammidesse läbi arvetekeskuse, aga see protsess on hetkel muutumas vabamaks [42]. Ilma arvetekeskuse lepinguta ei olnud hetkel võimalik töö autoril saata testimiseks e-arvet kasutusolevatesse raamatupidamise programmidesse. Selles hoolimata saadi valideerida e-arve korrektne parsimine, kasutades Envoice e-arve importimise võimalust. Joonisel 30 on kujutatud tulemus, mis oli positiivne: andmed olid korrektselt välja loetud ja samuti oli nähtav manuses olev tšeki pilt PDF failina.



Joonis 30. E-arve importimise kontrolli tulemus.

6 Järeldused

Järgnevalt antakse ülevaade töös saadud tulemustest ning ja tuuakse välja, mis kasu võiks loodud lahendusest olla ning arutletakse tulevikusuunade osas.

6.1 Loodud lahendus ärilises vaates

Tšekkide digitaliseerimiseks loodud lahendus andis etteantud sisendandmete (mille hulgas oli äärejuhtumeid) tšekkide täieliku digitaliseerimise tulemuseks 74% ja leitud andmete õigsuse tulemuseks 91%. See näitab, et täielikult seda protsessi automatiseerida võimalik ei olnud ning enne lõplikku kuludokumendi kinnitust on manuaalne kontroll siiski vajalik. Seetõttu oleks hea automaatsele protsessile lisada kinnituse etapp, millega kontrollitakse andmed (leitud andmed ja pilt oleks ühes vaates) ja ühe nupuvajutusega saadakse täielikult verifitseeritud e-arve.

Ajaliselt pakub loodud lahendus 94% ajavõitu manuaalse sisestuse ees (manuaalsel sisestuse keskmine ajakulu oli 2,1 minutit ja loodud lahendusel keskmiselt 8 sekundit), mis on väga märgatav kokkuhoid. Sellele võib küll lisanduda eelnevalt pakutud väike kontrollimise aeg.

Kokkuvõttena suudab antud lahendus hõlbustada 80% võrra raamatupidamises olevat kulutšekkide sisestuse protsessi.

Lisaks andmeväljade digitaliseerimisele pakub antud lahendus algdokumendi säilitamise võimalust. Töös selgitati välja, et pabertšeki digitaalne koopia on seaduslik ning lahenduses realiseeriti tšeki pildi lisamine e-arvele. Tänu sellisele lahendusele ei ole vajalik enam tšeki füüsiline transport raamatupidaja juurde, sellest koopia tegemine ning nende säilitamine füüsiliselt 7 aastat. Loodud lahendusega on algdokumendi koopia e-arve sees, mis saadetakse raamatupidamise tarkvarasse, kus ta koos teiste dokumentiga säilib 7 aastat. See on samuti märgatav protsessi lihtsustamine.

Sarnaselt raamatupidamise protsessile saaks kohandustega kasutada loodud lahendust kuluaruandluse protsessis, mis on suurtes ettevõtetes enne raamatupidamist ja seisneb korduvas andmete dubleerimises ja algdokumentide saatmise kuludes.

6.2 Loodud lahendus tehnilises vaates

Prototüübi valideerimise käigus selgus, et mõnel korral ei suudetud pildilt tausta eemaldada või jäi veidi tausta ääri ikkagi pildile, mis tekitas optilise tekstituvastuses müra. Seetõttu saaks pildi lõikamise viise veel täiendada. Töö käigus leiti, et liikuval pildil on äärte tuvastust palju lihtsam teha ning kui täis-lahenduse puhul realiseeritakse tšekkide pildistamiseks mobiilirakendus, siis oleks mõistlik teha äärte tuvastust juba pilti tehes (kasutaja näeb ja saab ise kinnitada ala õigsust). See võib anda veelgi parema tulemuse.

Samuti avastati, et optiline tekstituvastus ei suutnud alati õiget sõna välja lugeda ning tähed ja numbrid olid vahel sassi aetud. Selgus, et tšekil olev erinevas suuruses tekst ning lühendid mõjutavad tulemust. Siinkohal tasub mõelda optilise tekstituvastuse täiustamise peale, sest kasutatud optilise tekstituvastuse mootor Tesseract OCR võimaldab ka tuvastust ja keeli treenida, mis võib sellise sarnase temaatikaga piltide tuvastamisel abiks olla.

Kokkuvõttena tõestas prototüübi loomine, et soovitatud ja välja valitud tehnoloogiad sobisid selliseks ülesandeks ning loodud lahendus näitas häid tulemusi. Prototüüp tagastas täiesti valiidse e-arve, mida oli võimalik importida.

Loodud lahendust, selle osasid või isegi teguviise on võimalik kergelt taaskasutada. Optilist tekstituvastuse protsessi ja selle jaoks pildi eeltöötlus on kasulik ka muudele valdkondadele. Raamatupidamisalaste väärtuste tekstist leidmise protsess ja reeglid on disainitud eraldi osana ning seetõttu kergelt kasutatav ka muude sisendite korral. Näiteks võimalik proovida PDF-kujul oleva arve teisendamist e-arveks. Seega annab töös loodud lahendus võimaluse ka teisi raamatupidamise protsesse hõlbustada.

7 Kokkuvõte

7.1 Tehtud töö ja tulemused

Käesolev lõputöö seisnes ressursikulukate raamatupidamise tööprotsessi etappide automatiseerimises. Töö algul seati eesmärk oluliselt kiirendada ja parendada arvete sisestuse protsessi automatiseerimise teel.

Töös taheti teada, kui hästi on võimalik paberkandjal olevaid arveid viia digitaalsele kujule – kas seda on võimalik teha ilma inimest kaasamata ja millise eelise annaks see manuaalse protsessi ees.

Selleks uuriti põhjalikult ärilist tausta ning kaardistati olemasolevad lahendused, mis andis veelkord kinnitust probleemi aktuaalsusele. Raamatupidamisprotsesse hõlbustavaid lahendusi on viimasel ajal tänu e-arve tulekule tehtud mitmeid, kuid arve kohene digitaliseerimise osa neil hetkel puudub.

Samuti korraldati raamatupidajate vahel küsitlus. Küsitlus andis arvatud tagasisidet, et manuaalne protsess eksisteerib ning omad puuduseid. Pakutavat lahendust hindasid raamatupidajad hästi ning aitasid ärilise nõuga kogu töö vältel. Järgnevalt uuriti tehnilisi võimalusi ja sarnaseid töid ning tulemusena leiti sobivad tehnoloogiad. Täpsemalt plaaniti kasutada optilist tekstituvastust, reeglipõhist info eraldamist ning e-arve kasutamist.

Seejärel loodi prototüüp, milles keskenduti pabertšeki digitaliseerimisele. Prototüübi sisendiks on pildifail ning väljundiks e-arve (XML fail). Täpsemate sammudena tehakse prototüübis:

- pildi töötlus (suurimateks osadeks automaatne tausta eemaldus ja lävisegmentimine),
- optiline tekstituvastus (kasutades Tesseract OCR'i),

- reeglipõhine info eraldamine (regulaaravaldiste ja erineva tähtsusega reeglite loogika teel),
- saadud info valideerimine (vastavalt andmetüüpidele, formaatidele ja seostele),
- e-arve koostamine.

Töö valideerimise käigus katsetati prototüüpi väga erinevate testandmetega, milles võrreldi nii automaatselt kui ka manuaalselt erinevate prototüübi osade toimimist. Tähtsaima osana võrreldi tuvastamise protsenti ja ajakulu. Prototüüp suutis tuvastada täielikult ja õigesti 74% tšekkidest. Loodud lahendus protsessis tšekki keskmiselt 8 sekundit, mis teeb ajavõiduks manuaalse protsessi ees 94%. Lisaks tõestati, et loodud lahendus aitab ka algdokumendi säilitamise protsessi hõlbustada.

7.2 Püstitatud eesmärgid

Kas raamatupidamise tööprotsessi on võimalik oluliselt kiirendada ja parendada, kui automatiseerida arvete sisestamise etappi?

Töös selgitati välja, et raamatupidamise tööprotsessi on võimalik oluliselt kiirendada arvete sisestamist automatiseerides. Loodud lahendus võimaldas automaatset andmesisestust ning algdokumendi efektiivset kohustuslikku säilitamise protsessi.

Kas paberkandjal olevate tšekkide viimine digitaalsele kujule annab märgava eelise võrreldes manuaalse sisestusprotsessiga?

Loodud lahendus võimaldas 94% ajavõitu, seega paberkandjal olevate tšekkide viimine digitaalsele kujule annab märgatava eelise võrreldes manuaalse protsessiga.

Mis määral saab paberkandjal olevat tšekki viia digitaalsele kujule ilma inimest kaasamata (kas loodav lahendus suudaks väljastada sobiva digitaalse arve või peab inimene seda kontrollima ja täiendama)?

Paberkandjal olevate tšekkide digitaalsele kujule viimise protsessis on hetke loodud lahenduse järgi inimese osa vajalik. Selleks osaks on digitaalse arve kontrollimine ja vajadusel täiendamine, sest loodud lahendus suutis täielikult tuvastada 74% arvetest

ning andmete õigus oli 91%. Seega väike inimese osa on arvete digitaliseerimise hetkel loodud lahenduse juures vajalik.

7.3 Edasised võimalused

Hetkel loodud prototüüpi kasutusse võttes peaks lisama sellesse automaatsesse protsessi väike kontroll raamatupidaja poolt enne konteerimist, sest alati ei suudetud kõiki tšeki välja tuvastada. Selle osakaalu vähendamiseks või täiesti eemaldamiseks võiks proovida tekstituvastust treenida vastavalt tšeki peal oleva infole (lühendid ja mustrid). Samuti peaks ajapikku reegleid täiendama, mis on reeglisüsteemi puhul tavaline protsess ja annaks tuvastamisel parema tulemuse.

Käesoleva töö raames õpiti tundma suurel määral optilise tekstituvastuse ja pilditöötluse võimalusi. Nende põhjal saadi kinnitust, et pildi eelnev töötlemine on tähtis osa ning segava tausta peab saama võimalikult hästi eemaldatud. Seda võiks proovida täis-lahenduse korral täiustada pildistamise ajal tšeki ala tuvastusega.

Loodud lahendust saaks hõlpsasti integreerida juba olemasolevatesse arvete haldamise tarkvaradesse. Samuti hinnati loodud lahendus olemaks jätkusuutlik, sest lahendus oli disainitud eraldi osadena ning info eraldamist on võimalik kasutada näiteks muu teksti peal samuti (näiteks PDF-arve konvertimisel e-arveks). Lisaks saaks loodud lahendust kasutada ka veidi muudel eesmärkidel, näiteks suurte ettevõtete kuluaruandluse protsessi lihtsustamiseks.

7.4 Lõppsõna

Töö tulemusena veenduti, et raamatupidamises on võimalik manuaalset arvete sisestust infotehnoloogia vahenditega hõlbustada. Päris inimest kaasamata ei ole see hetkel võimalik. Esialgsele plaanitud protsessile oleks vajalik lisada väike kontrollimise etapp või proovida prototüüpi täiendada järelduses mainitud näpunäidetega. Aga sellest hoolimata annaks loodud lahendus juba praegusel hetkel märgatava ajavõidu ja hõlbustaks raamatupidajate tööd.

Kasutatud kirjandus

[1] Majanduslikult aktiivsed ettevõtted ja kasumitaotluseta üksused õigusliku vormi järgi, aasta. (2017) – Statistikaamet. http://www.stat.ee/68777 (07.05.2017)

[2] Eesti Riigiportaal, Raamatupidamine. https://www.eesti.ee/est/ettevotte_raamatupidamine (07.05.2017)

[3] Neudorf, R. (2010). Eestis valmib järjekordne e-lahendus: e-kviitung. – Postimees, Tehnika. http://tehnika.postimees.ee/335676/eestis-valmib-jarjekordne-e-lahendus-e-kviitung (07.05.017)

[4] Mets, M. (2011). Eesti uus suur IT-asi: e-kviitung. – Arvutimaailm, 7-8/11, 16-21.

[5] Tallinna Kaubamaja kontsern võttis kasutusele e-tšeki (2013). – Tallinna Kaubamaja Grupp. http://www.tkmgroup.ee/investor/tallinna-kaubamaja-kontsern-vottis-kasutusele-e-tseki (07.05.2017)

[6] Eerme, M. (2016). E-kviitung vähendab pabertšekkide osatähtsust Eesti kaubanduses. – Äripäev Kaubandus.ee. http://www.kaubandus.ee/uudised/2016/05/25/e-kviitung-vahendab-pabertsekkide-osatahtsust-eesti-kaubanduses (07.05.2017)

[7] Omniva e-kviitungi projekt jõudis Euroopa parimate hulka (2017). – Ituudised.ee. http://www.ituudised.ee/uudised/2017/01/09/omniva-e-kviitungi-projekt-joudis-euroopa-parimate-hulka (07.05.2017)

[8] Kviitung (beeta). https://www.kviitung.ee (07.05.2017)

[9] Johanson, A. (2016). Tähtaeg läheneb, kuid napilt pool avalikust sektorist on e-arvetele üle läinud. – Postimees, Majandus. http://majandus24.postimees.ee/3841555/tahtaeg-laheneb-kuid-napilt-pool-avalikust-sektorist-on-e-arvetele-ule-lainud (07.05.2017)

[10] Tsekk.ee. http://tsekk.ee/ (07.05.2017)

[11] Pau, A. (2017). Uus teenus Eesti turul: pildista tšekid mobiiliga ja laadi üles. – Postimees, Tehnika. http://tehnika.postimees.ee/4054499/uus-teenus-eesti-turul-pildista-tsekid-mobiiliga-ja-laadi-ules (07.05.2017)

[12] Sibold, G (2016). Tsekk.ee soovib muuta maailma paberkviitungitest priiks. – Geenius Meedia. https://geenius.ee/rubriik/nadala-idufirma/tsekk-ee-soovib-muuta-maailma-paberkviitungitest-priiks/ (07.05.2017)

[13] Envoice. https://envoice.eu (07.05.2017)

[14] Pau, A. (2016). Eesti idufirma aitab raamatupidajatel tšekimajandusega hakkama saada. – Postimees, Tehnika. https://tehnika.postimees.ee/3834901/eesti-idufirma-aitab-raamatupidajatel-tsekimajandusega-hakkama-saada (07.05.2017)

[15] Meres, R. (2015). Milline on Eesti keskmine raamatupidaja? – SimplBooks. http://www.simplbooks.ee/2015/11/milline-on-eesti-keskmine-raamatupidaja/ (07.05.2017)

[16] Raamatupidamise seadus. (2017) – Riigi Teataja. https://www.riigiteataja.ee/akt/127122016003 (07.05.2017)

[17] Rahandusministeerium. E-arved. http://www.fin.ee/e-arved (07.05.2017)

[18] Eesti Pangaliit. E-arve. http://www.pangaliit.ee/et/arveldused/e-arve (07.05.2017)

[19] Eesti Pangaliit. Eesti e-arve kirjeldus. http://www.pangaliit.ee/images/files/E-arve/Eesti_e-arve_kirjeldus_ver1.2_est.pdf (07.05.2017)

[20] Nshuti, C. N. (2015). Mobile Scanner and OCR (A first step towards receipt to spreadsheet).

https://pdfs.semanticscholar.org/f5cc/304b666f04096dd3b998f442a0f454b6e511.pdf (07.05.2017)

[21] Walker, D., Lund, W., Ringger, E.. (2010). Evaluating models of latent document semantics in the presence of OCR errors. - EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 240-250. http://dl.acm.org/citation.cfm?id=1870682 (07.05.2017)

[22] Mansfield, P., Levy, M. (2014). Identification of regions of a document. https://www.google.com/patents/US8832549 (07.05.2017)

[23] Nuggehalli, J. (2013). Expense Report System With Receipt Image Processing. https://www.google.com/patents/US20130232040 (07.05.2017)

[24] Sebastiani, F. (2002). Machine learning in automated text categorization. – ACM Computing Surveys, 34-1, 1-47. http://dl.acm.org/citation.cfm?id=505283 (07.05.2017)

[25] Dutta, A., Meilicke, C., Niepert, M., Ponzetto, S. (2013). Integrating open and closed information extraction: challenges and first steps. – NLP-DBPEDIA'13 Proceedings of the 2013th International Conference on NLP & DBpedia, 1064, 50-61. http://dl.acm.org/citation.cfm?id=2874485 (07.05.2017)

[26] Bannour, S., Audibert, L., Soldano, H. (2013). Ontology-based semantic annotation: an automatic hybrid rule-based method. – ACL 2013, 139-143. https://hal.archives-ouvertes.fr/hal-01074936/document (07.05.2017)

[27] Anno, T., Mizutani, M., Anno, G., Takada, S. (1994). Accounting book automatic entering device. https://www.google.com/patents/US5371680 (07.05.2017)

[28] Foth, T., Pierce, J., Citron, H., Wolfman, G., Romansky, B. (2003). Method and system for importing invoice data into accounting and payment programs. https://www.google.com/patents/US20030110128 (07.05.2017)

[29] Berry, C., Crowningshield, C., Marsh, D. (2004). Invoice entry. https://www.google.com/patents/US6832208 (07.05.2017)

[30] Benett, S. (2009). Invoice processing system. https://www.google.com/patents/US7509288 (07.05.2017)

[31] Euchner, J., Foth, T. (2008). Image based invoice payment with digital signature verification. https://www.google.com/patents/US20080147561 (07.05.2017)

[32] Larsson, A., Segeras, T. (2016) Automated invoice handling with machine learning and OCR. – TRITA-STH, 2016:53. http://www.diva-portal.org/smash/record.jsf?aq2=%5B%5B%5D%5D&af=%5B%5D&searchType=SIMPLE&language=en&pid=diva2%3A934351&aq=%5B%5B%5D%5D&sf=all&aqe=%5B%5D&sortOrder=author_sort_asc&onlyFullText=false&noOfRows=50&dspwid=-8950&dswid=4212 (07.05.2017)

[33] Boström, C., Herelius, J., Hugosson, M., Maleev, S. (2016). Automatic reading and interpretation of paper invoices: ADC invoice. – Independent Project in Computer and Information Engineering, 2016-014. http://www.diva-portal.org/smash/record.jsf?aq2=%5B%5B%5D%5D&af=%5B%5D&searchType=SIMPLE&language=en&pid=diva2%3A967971&aq=%5B%5B%5D%5D&sf=all&aqe=%5B%5D&sortOrder=author_sort_asc&onlyFullText=false&noOfRows=50&dspwid=7336&dswid=4212 (07.05.2017)

[34] Kaskman, R. Mobiilirakendus kulude jälgimiseks ostutšekkide skaneerimise funktsionaalsusega : magistritöö. Tallinna Tehnikaülikool, Tallinn, 2015. https://digi.lib.ttu.ee/i/?3554 (07.05.2017)

[35] Tesseract OCR. https://github.com/tesseract-ocr/tesseract/wiki (07.05.2017)

[36] OpenCV. http://opencv.org/ (07.05.2017)

[37] Tesseract OCR. Improving the quality of the output. https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality (07.05.2017)

[38] Wikipedia. Käibemaksukohustuslase number. https://et.wikipedia.org/wiki/K%C3%A4ibemaksukohustuslase_number (07.05.2017)

[39] Oracle. Java Architecture for XML Binding (JAXB). http://www.oracle.com/technetwork/articles/javase/index-140168.html (07.05.2017)

[40] iText. Java PDF library. http://itextpdf.com/ (07.05.2017)

[41] XML Validator - XSD (XML Schema). http://www.freeformatter.com/xml-validator-xsd.html (07.05.2017)

[42] Veskioja, I. (2017). E-arvete vahendamiselt enam raha teenida ei saa. – Envoice. https://envoice.eu/et/e-arvete-vahendamiselt-enam-raha-teenida-ei-saa/ (07.05.2017)

Lisa 1 – Küsitluse ankeet raamatupidajate vahel

1. Mitu paberkandjal olevat kviitungit/tšekki Te keskmiselt iga nädal raamatupidamistarkvarasse sisestate?*

2. Kui palju kulub keskmiselt aega ühe sellise kuludokumendi sisestamise peale?* Vastuses palun täpsustada umbkaudne aeg sekundites või minutites.

3. Kui suure osa moodustavad e-arved kogu kuludokumentidest, mida sisestate raamatupidamistarkvarasse?*
   Vastusega sooviks näha, mis määral kasutatakse kuludokumentide haldamisel elektroonilisi arveid ja nende importimist.

4. Kas olete kuulnud paberkandjal olevate kuludokumentide sisestuse automatiseerimisest? *

Plaanitav lahendus kuludokumentide sisestuse lihtsustamiseks

- Telefoniga pildistatakse kviitungit/tšekki (näiteks ettevõtte omanik ise)
- Tarkvara konverteerib pildi e-arveks (kus on kogu kviitungi/tšeki info juba täidetud)
- E-arve (kus küljes on ka esialgne pilt) on raamatupidajale kättesaadav
- Raamatupidaja impordib e-arve raamatupidamistarkvarasse

Oodatavad plussid:

- raamatupidaja võidab ajas - ei pea enam manuaalselt sisestama kviitungil/tšekil olevat infot
- kviitungi/tšeki koopia säilitatakse e-arve küljes (ei ole arvatavasti vaja eraldi paberil koopiat teha)

Täpsustus: lõputöös ei realiseerita kogu lahendit, vaid leitakse sobilik lahendusviis tarkvara tuumiku jaoks, mis tõestaks üldise lahenduse toimimist.

5. Mida arvate plaanitavast lahendusest? Milliseid puudujääke näete? Soovitusi?*

Lisa 2 – Prototüübi lähtekood

Prototüübi lähtekood asub versioonihaldussüsteemis järgneval lingil:
https://github.com/r1ek/ReceiptDigitalizer

Lisa 3 – Genereeritud e-arve XML formaadis

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<E_Invoice>
    <Header>
        <Date>2017-05-01</Date>
        <FileId>11199045_00000002320</FileId>
        <Version>1.2</Version>
    </Header>
    <Invoice invoiceId="00000002320000022127" regNumber="12972725"
sellerRegnumber="11199045">
        <InvoiceParties>
            <SellerParty>
                <Name>Klick Eesti AS</Name>
                <RegNumber>11199045</RegNumber>
            </SellerParty>
            <BuyerParty>
                <Name>Lagom OÜ</Name>
                <RegNumber>12972725</RegNumber>
            </BuyerParty>
        </InvoiceParties>
        <InvoiceInformation>
            <Type type="DEB"/>
            <DocumentName>Receipt</DocumentName>
            <InvoiceNumber>00000002320000022127</InvoiceNumber>
            <InvoiceDate>0017-03-17</InvoiceDate>
        </InvoiceInformation>
        <InvoiceSumGroup>
            <TotalVATSum>15.00</TotalVATSum>
            <TotalSum>89.99</TotalSum>
            <Currency>EUR</Currency>
        </InvoiceSumGroup>
        <InvoiceItem>
            <InvoiceItemGroup>
                <ItemEntry>
                    <Description>EXPENSE REPORT</Description>
                    <ItemSum>74.99</ItemSum>
                    <VAT>
                        <VATRate>20</VATRate>
                        <VATSum>15.00</VATSum>
                    </VAT>
                    <ItemTotal>89.99</ItemTotal>
                </ItemEntry>
            </InvoiceItemGroup>
        </InvoiceItem>
        <AttachmentFile>
            <FileName>tsekk (83)processed.png.pdf</FileName>

<FileBase64>JVBERi0xLjQ...CROPPED_THIS_LONG_PART_FOR_WORD_DOCUMENT...D
g3NQolJUVPRgo=</FileBase64>
        </AttachmentFile>
        <PaymentInfo>
            <Currency>EUR</Currency>

<PaymentDescription>Receipt00000002320000022127</PaymentDescription>
            <Payable>NO</Payable>
            <PaymentTotalSum>89.99</PaymentTotalSum>
            <PayerName>Lagom OÜ</PayerName>
            <PaymentId>00000002320000022127</PaymentId>
```

```
            <PayToAccount></PayToAccount>
            <PayToName>Klick Eesti AS</PayToName>
        </PaymentInfo>
    </Invoice>
    <Footer>
        <TotalNumberInvoices>1</TotalNumberInvoices>
        <TotalAmount>89.99</TotalAmount>
    </Footer>
</E_Invoice>
```

Lisa 4 – Statistika aruanne

```
===STATISTICS===
(1) processing time: 10.027s
(2) processing time: 6.606s
(3) processing time: 11.759s
(4) processing time: 7.714s
(5) processing time: 7.334s
(6) processing time: 6.643s
(7) processing time: 13.461s
(8) processing time: 24.905s
(9) processing time: 11.721s
(10) processing time: 11.097s
(11) processing time: 5.619s
(12) processing time: 6.69s
(13) processing time: 5.119s
(14) processing time: 6.801s
(15) processing time: 5.738s
(16) processing time: 7.968s
(17) processing time: 7.942s
(18) processing time: 10.843s
(19) processing time: 12.926s
(20) processing time: 10.073s
(21) processing time: 9.279s
(22) processing time: 5.517s
(23) processing time: 4.915s
(24) processing time: 7.18s
(25) processing time: 5.531s
(26) processing time: 10.919s
(27) processing time: 9.779s
(28) processing time: 6.659s
(29) processing time: 8.106s
(30) processing time: 11.531s
(31) processing time: 6.652s
(32) processing time: 7.79s
(33) processing time: 13.38s
(34) processing time: 10.078s
(35) processing time: 10.564s
(36) processing time: 14.614s
(37) processing time: 17.481s
(38) processing time: 4.929s
(39) processing time: 11.076s
(40) processing time: 8.895s
(41) processing time: 11.867s
(42) processing time: 8.76s
(43) processing time: 7.154s
(44) processing time: 10.304s
(45) processing time: 6.258s
(46) processing time: 5.803s
```

```
(47) processing time: 3.847s
(48) processing time: 6.887s
(49) processing time: 5.467s
(50) processing time: 7.503s
(51) processing time: 6.096s
(52) processing time: 6.434s
(53) processing time: 8.274s
(54) processing time: 21.921s
(55) processing time: 2.707s
(56) processing time: 4.999s
(57) processing time: 5.171s
(58) processing time: 24.81s
(59) processing time: 2.119s
(60) processing time: 4.694s
(61) processing time: 5.587s
(62) processing time: 6.411s
(63) processing time: 4.595s
(64) processing time: 8.082s
(65) processing time: 4.41s
(66) processing time: 7.488s
(67) processing time: 11.46s
(68) processing time: 10.982s
(69) processing time: 46.146s
(70) processing time: 9.31s
(71) processing time: 6.687s
(72) processing time: 6.809s
(73) processing time: 6.532s
(74) processing time: 9.947s
(75) processing time: 5.703s
(76) processing time: 5.758s
(77) processing time: 3.736s
(78) processing time: 4.99s
(79) processing time: 4.812s
(80) processing time: 4.53s
(81) processing time: 6.596s
(82) processing time: 5.332s
(83) processing time: 9.391s
(84) processing time: 7.694s
(85) processing time: 13.303s
(86) processing time: 8.85s
(87) processing time: 5.154s
(88) processing time: 3.752s
(89) processing time: 3.857s
(90) processing time: 5.036s
 ----------------------------
average execution time: 8s

===STATISTICS===
nothingFOUND:        0 / 90
allFOUND:            7 / 90
sellerRegOrVatFOUND: 83 / 90
```

```
twoAmount_okCalc:    69 / 90
allAmountsFOUND:     33 / 90
sellerNameFOUND:     42 / 90
dateFOUND:           68 / 90
idFOUND:             67 / 90
totalAmountFOUND:    81 / 90
vatRateFOUND:        70 / 90
vatAmountFOUND:      40 / 90
currencyFOUND:       65 / 90
all:                  90
```