УНИВЕРЗИТЕТ У БАЊОЈ ЛУЦИ
ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Љубиша Црнадак

НЕНАДГЛЕДАНА КЛАСИФИКАЦИЈА РУКОМ ПИСАНИХ РИЈЕЧИ

дипломски рад

Бања Лука, април 2011.

Тема: **НЕНАДГЛЕДАНА КЛАСИФИКАЦИЈА РУКОМ ПИСАНИХ РИЈЕЧИ**

Кључне ријечи:
Препознавање облика
Препознавање руком писаног текста
UPGMA алгоритам

Комисија: **ред. проф. др Милорад Божић, предсједник**
ванр. проф. др Зденка Бабић, ментор
мр Владимир Рисојевић, члан

Кандидат:
Љубиша Црнадак

УНИВЕРЗИТЕТ У БАЊОЈ ЛУЦИ
ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ
КАТЕДРА ЗА ОПШТУ ЕЛЕКТРОТЕХНИКУ

Предмет: Дигитална обрада слике

Тема: НЕНАДГЛЕДАНА КЛАСИФИКАЦИЈА РУКОМ ПИСАНИХ РИЈЕЧИ

Задатак: Упознати се са проблемом ненадгледане класификације руком писаних ријечи и могућим примјенама. Предложити алгоритам за ненадгледану класификацију руком писаних ријечи издвојених из скенираних докумената. Имплементирати предложени алгоритам и практично потврдити оправданост поступка.

Ментор: др Зденка Бабић, ванредни професор

Кандидат: Љубиша Црнадак (1246/10)

Бања Лука, април 2011.

мојим родитељима

Садржај

Уз рад је приложен CD.

Списак слика

Глава 1

УВОД

Препознавање и анализа дигитализованих докумената представља важну подобласт научне дисциплине препознавања облика. Појам анализе дигитализованих докумената, у ужем смислу, подразумијева анализу распореда елемената унутар документа и њихових атрибута као што су величина, типографија и сл.

Појам препознавања дигитализованих докумената, подразумијева успостављање коресподенције између слике дигитализованог документа и електронске репрезентације, прије свега текстуалног дијела садржаја документа.

Јасно је да се ова два аспекта разумијевања садржаја дигитализованих докумената прожимају и да се најчешће посматрају здружено. У овом раду наглашено ће се посматрати појам препознавања текста, и то руком писаног текста, али не заборављајући шири аспект проблема препознавања и анализе дигитализованих докумената.

1.1 Препознавање: током писања - након писања

Да се не би створила забуна кратко ће бити појашњена основна подјела приликом дефинисања проблема препознавања текста писаног руком. Препознавање може да се врши током писања (нпр. оловком по површини осјетљивој на додир), а ова класа проблема се назива препознавање током писања, односно on-line препознавање. Други случај представља препознавање текста који је већ написан, затим неком фотографском методом дигитализован, тако да је у рачунару представљен као матрица тачака. Ова класа проблема се назива препознавање након писања, односно off-line препознавање и оно је тема овог рада.

Потребно је нагласити да су потпуно различити приступи приликом рјешавања ове двије врсте проблема. Ово је потпуно разумљиво због тога што подаци који се обрађују имају квалитативно веома различит карактер. На примјер, код on-line класе проблема као улазни параметар имплицитно фигурише вријеме, односно трендови помјерања оловке у простору (трајекторија), што код off-line метода једноставно није доступна информација.

Рад Plamondon-a и Srihari-a „On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey" је веома опсежна студија препознавања руком писаног текста, са преко 250 библиографских јединица [5].

У наставку рада, ради концизности ће се под појмом текста у дигитализованом документу подразумијевати текст писан руком.

1.2 Уочавање кључних ријечи

Појам препознавања текста је описан као поступак успостављања коресподенције између дијела слике дигитализованог документа и електронске репрезентације текстуалног садржаја документа. На препознавање се може гледати као на коначан циљ, у смислу жеље да се у потпуности успостави наведена коресподенција. Пошто се показало да је то велик изазов код текста писаног руком, тражен је начин да се остваре нешто скромнији циљеви које у ширем смислу и даље сматрамо као рјешења препознавања у одређеним ситуацијама и за одређену намјену. Један од таквих скромнијих циљева је претрага кључних ријечи у тексту, а техника је позната као „уочавање ријечи" (Wordspotting) .

Ова техника подразумијева претрагу унапријед дефинисане ријечи унутар дигитализованог документа. Практично је проблем препознавања постављен инверзно и почиње са информацијом која се тражи а не од документа који се анализира. Теоретски гледано, потпуно рјешење уочавања ријечи би значило ријешен проблем препознавања, али практично је јасно да је немогуће на овај начин доћи до коначног циља препознавања. Ипак, овај скромнији циљ има велик практични значај и представља покушај да се омогући претрага и проналажење дигитализованих докумената који садрже одређену кључну ријеч.

1.3 Препознавање и класификација без претходног знања

Тренутна научна достигнућа за рјешавање проблема уочавања ријечи су углавном фокусирана у сврху претраге историјских докумената. Пошто је ријеч о одређеним регистрима архивске грађе за коју се пројектују, углавном се дио те грађе ручно обиљежи и користи за обучавање система препознавања. У овом раду посматраће се нешто другачији сценарио употребе уочавања ријечи, који не узима у обзир знање о документима који се анализирају, а назива се ненадгледана класификација.

1.4 Концепција рада

Домен примјене и дефиниција проблема који се жели ријешити ће бити описани у наредној глави.

Глава 3 - Описује поступке издвајања обиљежја која ће се користити за класификацију руком писаних ријечи.

Глава 4 - Представља поступак кластеризације који за циљ има класификацију ријечи са истим семантичким садржајем, са минималном грешком.

Глава 5 - Описује практичну реализацију и интерпретацију резултата остварених овим радом.

Посљедња глава даје тумачење добијених резултата у ширем контексту, у виду закључка.

Глава 2

ФОРМУЛАЦИЈА ПРОБЛЕМА

2.1 Домен примјене

Прије формалне формулације проблема, биће описан домен примјене, дефинисаће се циљеви који се желе постићи са практичног аспекта. Резултат овог рада би требало да омогући бржи и тачнији унос података, који су руком написани на папиру. Подаци који се желе превести у електронски облик би требало да се налазе на тачно одређеним мјестима на папиру, гдје је на одређен начин предвиђен и означен простор за првобитни унос података руком.

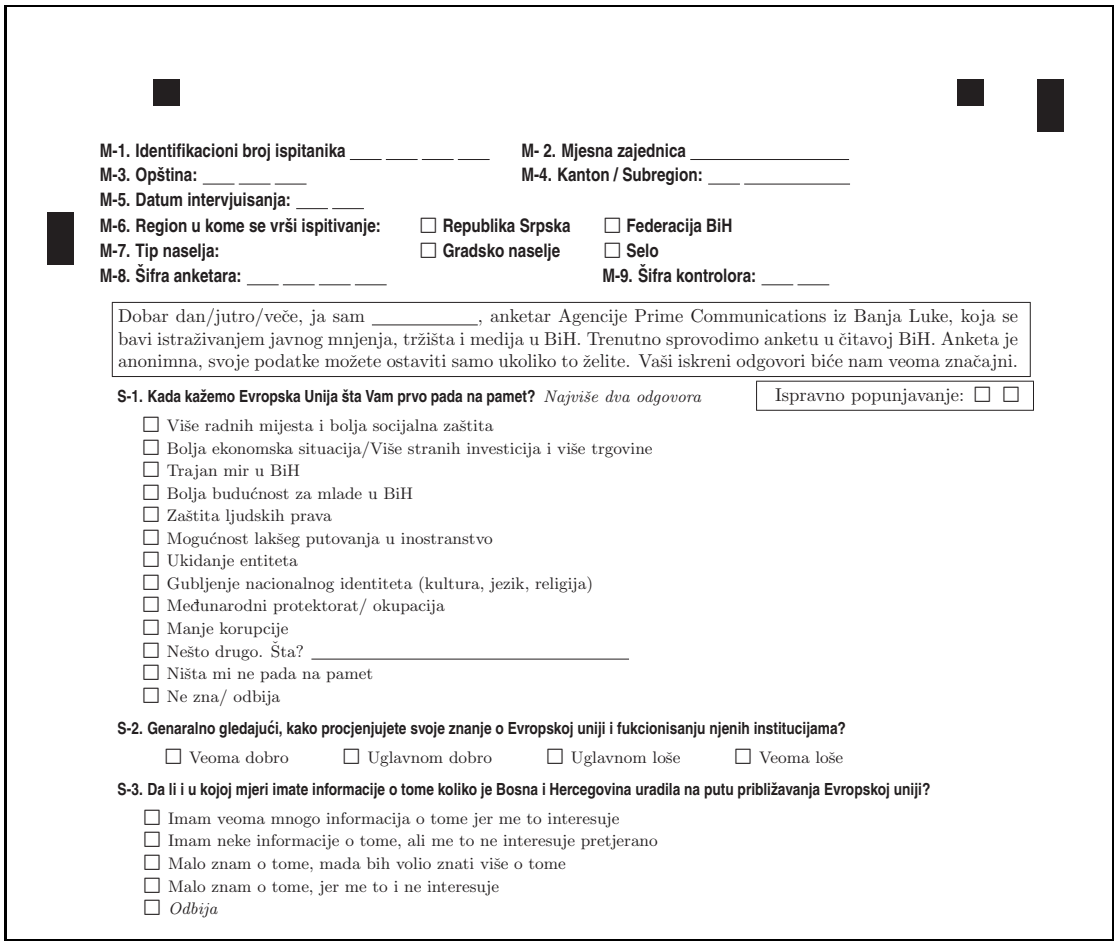
2.1.1 Врсте докумената за обраду

Документи који имају унапријед дефинисан шаблон су разни упитници, анкетни листови и остали табеларни шаблони који се користе приликом анкетирања, у научно-истраживачким пројектима, за потребе статистичке обраде и сл. На Слици 2.1 је приказан дио анкете која је обрађивана аутоматизованим системом, у којој постоје текстуална поља која се попуњавају руком. Ради се о истраживању свијести грађана о Европској унији, које је спроводила агенција за истраживање јавног мнијења Prime communications из Бањалуке. Након попуњавања жељених поља, скенирају се упитници и добију црно-бијеле слике сваке странице, за сваки упитник. Примјер ове фазе за други анкетни лист је дат на Слици 2.2, за истраживање које је спроводила агенција за маркетниг, пословни консалтинг и истраживање D2D Advertising из Бањалуке.

Код аутоматизованог система, на упитницима се дефинишу маркери који помажу при локализацији дијелова упитника који треба да се препознају, а најчешће се налазе на маргинама упитника. Претпостављајући да је сам упитник генерисан од стране истог аутоматизован система, јасно је да се на основу маркера могу лоцирати сви елементи упитника након скенирања, а самим тим и мјеста која би требало да садрже руком писан текст. На овај начин је практично ријешен проблем локализације, и већим дијелом је ријешен и проблем сегментације текста који се жели препознати.

Крајњи циљ овог аутоматизованог система је да се сви подаци са упитника преведу у електронски облик укључујући и елементе руком писаног текста. Узимајући у обзир да је ово и даље веома тежак задатак, у ову сврху ће се користити wordspotting и то на начин да се класификују уноси (узорци) са истим текстом у засебну групу, затим да се оператеру прикаже та група и на крају да се ручно означи о којем тексту је ријеч.

Дакле, жеља је да се направи аутоматизован систем којим би се текстуални дијелови упитника, односно анкетних листова, могли на много бржи и тачнији начин уносити,



M-1. Identifikacioni broj ispitanika ___ ___ ___ ___ M- 2. Mjesna zajednica _____
M-3. Opština: ___ ___ ___ M-4. Kanton / Subregion: ___ _____
M-5. Datum intervjuisanja: ___ ___
M-6. Region u kome se vrši ispitivanje: ☐ Republika Srpska ☐ Federacija BiH
M-7. Tip naselja: ☐ Gradsko naselje ☐ Selo
M-8. Šifra anketara: ___ ___ ___ ___ M-9. Šifra kontrolora: ___ ___

Dobar dan/jutro/veče, ja sam _____, anketar Agencije Prime Communications iz Banja Luke, koja se bavi istraživanjem javnog mnjenja, tržišta i medija u BiH. Trenutno sprovodimo anketu u čitavoj BiH. Anketa je anonimna, svoje podatke možete ostaviti samo ukoliko to želite. Vaši iskreni odgovori biće nam veoma značajni.

S-1. Kada kažemo Evropska Unija šta Vam prvo pada na pamet? *Najviše dva odgovora* Ispravno popunjavanje: ☐ ☐

- ☐ Više radnih mijesta i bolja socijalna zaštita
- ☐ Bolja ekonomska situacija/Više stranih investicija i više trgovine
- ☐ Trajan mir u BiH
- ☐ Bolja budućnost za mlade u BiH
- ☐ Zaštita ljudskih prava
- ☐ Mogućnost lakšeg putovanja u inostranstvo
- ☐ Ukidanje entiteta
- ☐ Gubljenje nacionalnog identiteta (kultura, jezik, religija)
- ☐ Međunarodni protektorat/ okupacija
- ☐ Manje korupcije
- ☐ Nešto drugo. Šta? _____
- ☐ Ništa mi ne pada na pamet
- ☐ Ne zna/ odbija

S-2. Genaralno gledajući, kako procjenjujete svoje znanje o Evropskoj uniji i fukcionisanju njenih institucijama?

☐ Veoma dobro ☐ Uglavnom dobro ☐ Uglavnom loše ☐ Veoma loše

S-3. Da li i u kojoj mjeri imate informacije o tome koliko je Bosna i Hercegovina uradila na putu približavanja Evropskoj uniji?

- ☐ Imam veoma mnogo informacija o tome jer me to interesuje
- ☐ Imam neke informacije o tome, ali me to ne interesuje pretjerano
- ☐ Malo znam o tome, mada bih volio znati više o tome
- ☐ Malo znam o tome, jer me to i ne interesuje
- ☐ *Odbija*

Слика 2.1: Оригинални изглед дијела анкетног листа (са дозволом)

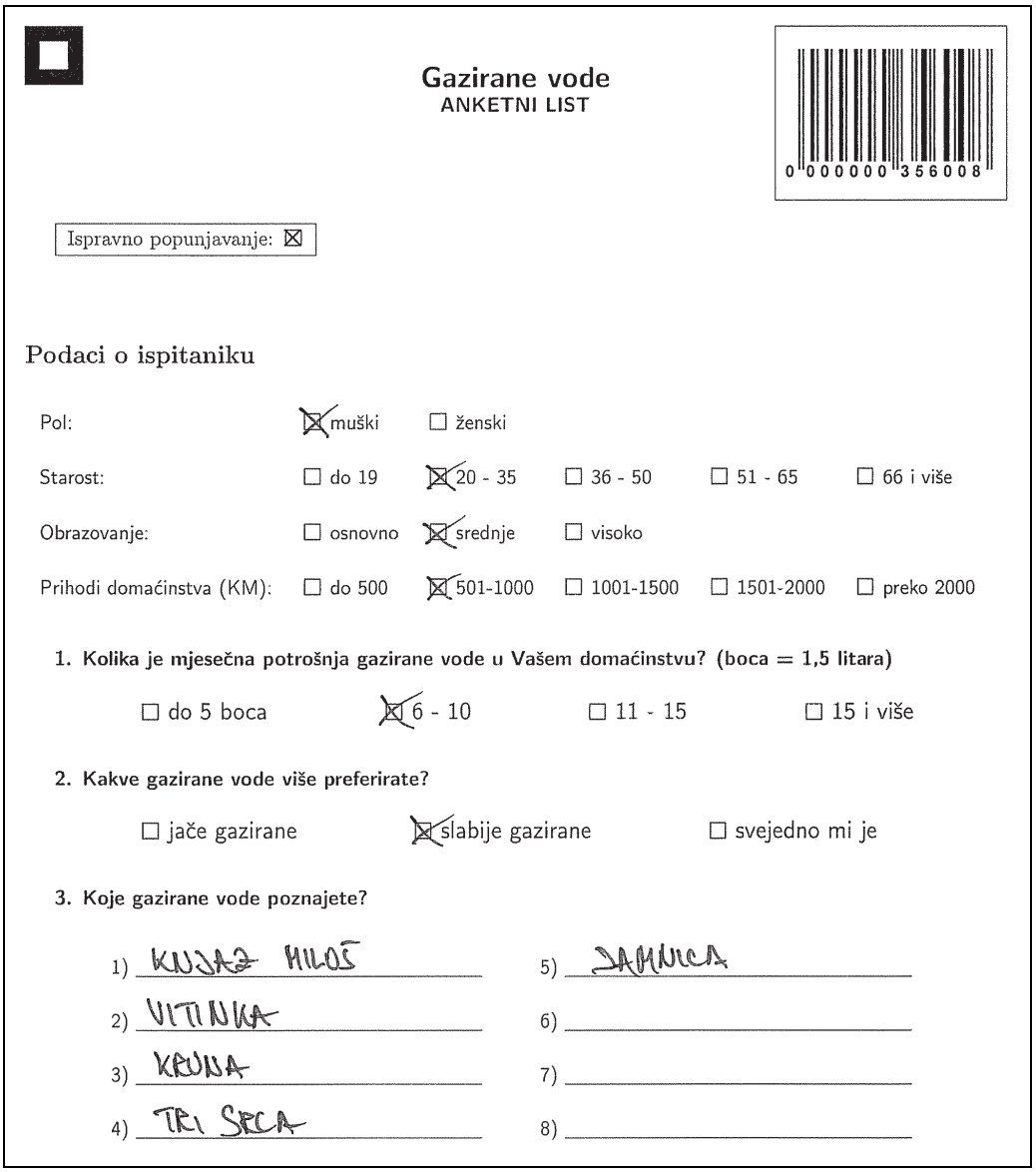
што би представљало један модул аутоматизованог система за унос цијелог упитника.

Овакав модул би посебно био погодан за унос података чији одговори чине релативно мален, али унапријед непознат скуп. На примјер, туристичка агенција би у својој анкети поставила питање: „У који град бисте најрадије отпутовали?". За статистичку обраду би се тражио списак свих градова, у виду шифрарника, који су били присутни у одговорима.

2.2 Домен проблема

Потребно је истакнути да ће овај рад подразумијевати да је све анкете руком испунила једна особа - анкетар. Након скенирања (дигитализације) попуњеног упитника, реализује се локализација и сегментација дијелова упитника за које се сматра да садрже руком писан текст. Проблем овог рада се формулише након овог корака. Процес локализације и сегментације није интерес овог рада и неће бити детаљно приказан.

Дакле, сматраће се да постоје сегменти који су локализовани, над којима је потребно урадити предобраду у смислу основних али и нешто сложенијих морфолошких операција као што је уклањање линије на којој лежи текст писан руком и сл. Практично, те



Gazirane vode
ANKETNI LIST

0 0000000 356008

Ispravno popunjavanje: ☒

Podaci o ispitaniku

Pol:	☒ muški	☐ ženski			
Starost:	☐ do 19	☒ 20 - 35	☐ 36 - 50	☐ 51 - 65	☐ 66 i više
Obrazovanje:	☐ osnovno	☒ srednje	☐ visoko		
Prihodi domaćinstva (KM):	☐ do 500	☒ 501-1000	☐ 1001-1500	☐ 1501-2000	☐ preko 2000

1. Kolika je mjesečna potrošnja gazirane vode u Vašem domaćinstvu? (boca = 1,5 litara)

☐ do 5 boca ☒ 6 - 10 ☐ 11 - 15 ☐ 15 i više

2. Kakve gazirane vode više preferirate?

☐ jače gazirane ☒ slabije gazirane ☐ svejedno mi je

3. Koje gazirane vode poznajete?

1) KUJAČ MILOŠ
2) VITINKA
3) KRUNA
4) TRI SRCA
5) JAMNICA
6)
7)
8)

Слика 2.2: Скениран узорак дијела попуњеног анкетног листа (са дозволом)

операције треба да поправе квалитет скенираног текста, али и да уклоне све елементе шаблона који је одштампан прије било каквог уноса, а појављује се након скенирања попуњеног упитника у локализованим сегментима. У наставку ће бити више ријечи о систему за унос и предобраду података.

Претпоставља се да је текст у локализованим сегментима исте природе (на примјер, у питању под редним бројем три, на Слици 2.2 налази се осам поља у која се уноси текст исте природе - назив минералне воде). Прије свега, узима се иста локација на упитнику за све реализације (различито попуњене упитнике). Поред тога, узимају се и остале локације које садрже податке исте природе, водећи рачуна о позицији са које је екстрахован податак.

2.2.1 Дигитализација и обрада узорка

Претварање испуњеног упитника са папира у дигитални облик је први корак који је потребно да се уради. У практичном систему који је реализован коришћена је техника црног-бијелог скенирања са резолуцијом од 300 тачака по инчу (300 DPI).

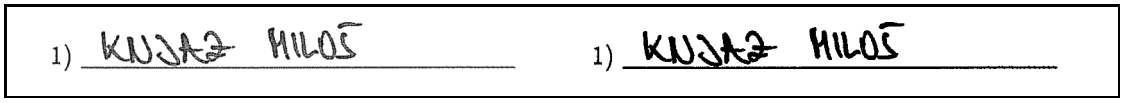
Ово се показало као минимална резолуција (у црно-бијелој техници) са којом је могуће остварити велику поузданост за читав систем (препознавање осталих елемена на упитнику, поред текста).

Након скенирања; фајлови се снимају коришћењем стандарда CCITT-G4 који је усвојио Comité Consultatif International Téléphonique et Télégraphique, данас познат као International Telecommunication Union. Овим стандардном, који се такође користи при трансмисији факсимила, постиже се велика компресија података без губитка информација.

2.2.1.1 Основне морфолошке операције

Након скенирања потребно је урадити основне морфолошке операције над скенираном узорком да би се уклонили непотребни дијелови и појачали (испунили) дијелови слике који су изостављени самим процесом дигитализације.

Прије свега се користе поступци дилатације и ерозије, затим се уклањају морфолошки елементи који имају површину испод неког прага, пошто на њих гледамо као на шум. Ово се посебно односи на дијелове упитника чији садржај је унапријед познат, и



Слика 2.3: Узорак прије и послије основне морфолошке обраде

за које се са сигурношћу зна да не би требало да садрже такве елементе. (нпр. дијелови упитника на којима су дефинисани маркери).

На Слици 2.3 је приказан дио руком писаног текста прије и послије основне морфолошке обраде. Ова обрада омогућава да се уклоне недостаци, настали усљед скенирања, али евентуално и неки други недостаци који су могли да оштете узорак прије самог скенирања.

2.2.1.2 Обрада узорка

Након основних морфолошких операција потребно је локализовати маркере који се налазе на скенираном упитнику, затим поредећи добијене податке са предефинисаним моделом, израчунати геометријске трансформације који пресликавају домен дигитализованог упитника у домен модела.

Локализација се врши тако што се узима релативно велик регион за који се са сигурношћу зна да се у њему налази маркер. Након приписивања лабеле сваком елементу у датом региону, за сваки елемент се рачунају морфолошки параметри (нпр. површина, облик и сл.). Одбацују се сви елементи који не задовоље одређене морфолошке услове за наведене параметре. Како су маркери веома специфичног геометријског облика и распореда, на овај начин се одбацују сви елементи осим маркера.

Након овог корака, потребно је идентификовати карактеристичне тачке маркера и упарити са тачкама модела маркера. Овај поступак се може урадити на два начина. Један је да се траже карактеристичне тачке на самом маркеру и очитају њихове координате, а други начин је да се на основу читавог скенираног маркера интерполира одговарајући геометријски модел маркера, те да се очитају координате тачака тако добијеног модела.

На крају, након идентификације одговарајућих тачака маркера, рачуна се здружена геометријска трансформација за све маркере који се налазе на једној страници, односно за читав упитник. Поставља се ограничење да је ријеч о трансформацији која може да ротира, транслира и пропорционално увећа координате идентификованих маркера да би се подударале са моделом. Оваква трансформација је у литератури позната као „крута трансформација" (енг. rigid transformation) . Да би тачност наведеног пресликавања била већа, користи се предетерминисан скуп тачака (теоретски су довољне само три тачке), а трансформација се рачуна минимизацијом средњеквадратне грешке пресликаних тачака и тачака модела.

Након тога, трансформација се примјени на читаву слику странице дигитализованог упитника. На овај начин слика је у потпуности преведена у координате модела и довољно је само очитати позиције жељених поља за обраду, односно на тај начин је ријешен проблем сегментације.

2.2.2 Издвајање обиљежја

Након локализације и морфолошке предобраде, издвајају се обиљежја која ће служити за класификацију ријечи.

Аутор је практично испитао два потпуно различита приступа, који се суштински разликују у карактеру одређивања обиљежја. Један је заснован на раду David Lowe-a, односно SIFT дескриптору [2]. Други је заснован на профилима који се екстрахују из дигитализованих слика ријечи [7], [8]. Како се други приступ показао супериорнијим, управо тај приступ је детаљно описан у овом раду.

Израчуната обиљежја се даље пореде методом динамичког увијања времена (енг. Dynamic Time Warping - DTW), која своје коријене и највећу примјену налази у области препознавања говора [11]. Након поређења парова узорака формира се матрица дистанци која показује разлике (или сличности) између било која два узорка.

2.2.3 Кластеризација

На крају је потребно извршити кластеризацију матрице дистанци да би се добили кластери за које кажемо да представљају рјешење проблема. Идеално рјешење би у засебне кластере издвојило слике ријечи на којима је исписана иста ријеч. За овај поступак се користи једна од стратегија агломеративне хијерархијске кластеризације која ће детаљно бити описана у наставку рада.

Глава 3

ИЗДВАЈАЊЕ ОБИЉЕЖЈА

3.1 Увод

Након сегментације, морфолошке обраде и припреме узорака потребно је издвојити карактеристична обиљежја којим ће бити описане сегментиране ријечи. Приликом избора обиљежја, пожељно је одабрати обиљежја тако да приликом поређења буду инваријантна на величину, као и на разне деформације до којих може доћи у претходним корацима издвајања ријечи за обраду. Постоји велика лепеза обиљежја која су препоручена за препознавање и обраду руком писаног текста, чији преглед се може пронаћи у литератури [4]. Аутор је за потребу рада испробао неколико тополошких обиљежја препоручених у [4]. Ипак, показало се да резултати нису били задовољавајући, тако да та обиљежја ипак нису у најбољој мјери осликавала проблем који овај рад покушава да ријеши.

За потребе wordspotting-а најчешће се користе двије врсте обиљежја. Једна врста се базира на SIFT методи [2]. Аутор је испробао обиљежја базирана на SIFT методи. Показало се да ова врста обиљежја није дала задовољавајуће резултате и поред тврдњи [9], тако да детаљније није обрађивана. Инфериорност SIFT методе наводе и други аутори [6].

Другу врсту обиљежја за потребе wordspotting-а представљају обиљежја базирана на профилима ријечи која се описује. Обиљежја базирана на профилима се користе у овом раду и практично се показало да дају задовољавајуће резултате.

3.2 Обиљежја базирана на профилима

Коришћење профила као основних обиљежја за успостављање сличности између дигитализованих, руком писаних ријечи је познато у литератури [6], [7], [8]. Заједничка особина ових обиљежја је да се за сваку колону слике за коју се одређује обиљежје добије једна скаларна вриједност. Нормализација се врши висином слике у пикселима ако другачије није назначено. Да би се визуелизовали профили који ће бити објашњени у наставку користиће се једна слика из узорка који је практично тестиран, а приказана је на Слици 3.1.

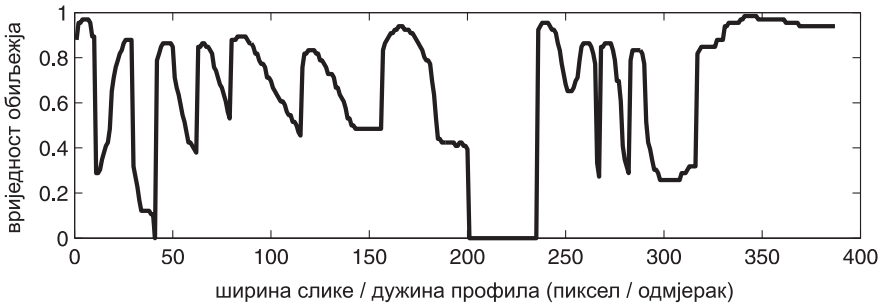
У наставку ће се усвојити конвенција да бијели пиксели представљају објекат, а црни пиксели представљају позадину. Сва обиљежја која ће се посматрати претпостављају црно-бијеле слике, без нијанси сиве. Објекат у овом случају представља руком написану ријеч и евентуално дијелове неких других објеката, ако је у процесу сегментације дошло

Слика 3.1: Један узорак из колекције која је коришћена у практичном дијелу рада

до грешке.

3.2.1 Горњи профил

Горњи профил се рачуна тако што се за сваку колону слике мјери удаљеност бијелог пиксела (најближег горњој ивици слике) од доње ивице слике. Под удаљеношћу у овом случају сматрамо број пиксела. Колона слике која не садржи пикселе бијеле боје може да представља размак између слова или ријечи. У овом случају профилу додјељујемо вриједност нула. Често се на овим мјестима рачуна просјечна вриједност



Слика 3.2: Горњи профил узорка са слике 3.1

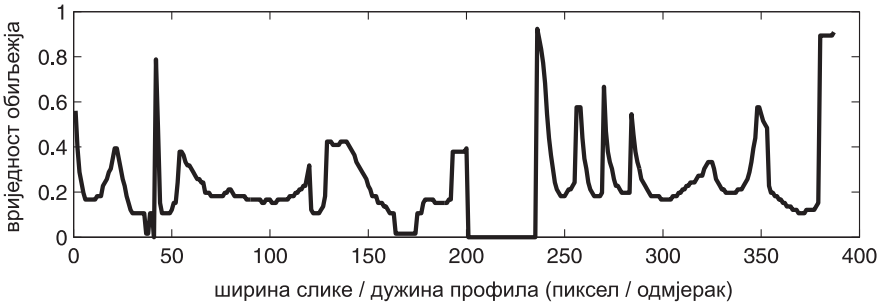
сусједних колона која има пикселе бијеле боје, али у практичном дијелу овог рада то није коришћено. Изглед горњег профила за Слику 3.1 је приказан на Слици 3.2.

3.2.2 Доњи профил

Доњи профил се рачуна аналогно горњем профилу, при чему се удаљеност мјери од доње ивице слике, за бијели пиксел најближи доњој ивици слике. Изглед доњег профила је приказан на Слици 3.3.

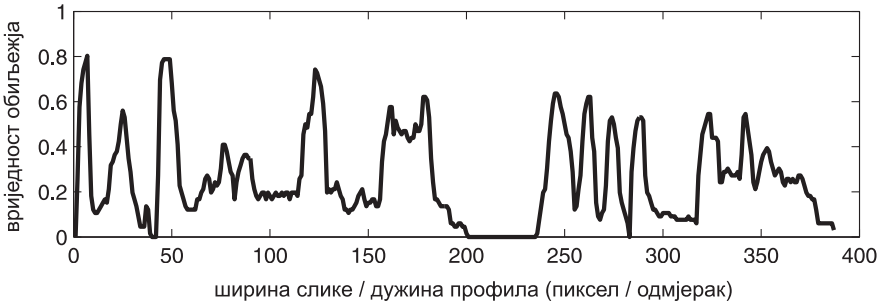
3.2.3 Профил пројекције

Профил пројекције се рачуна тако што што се за сваку колону слике сабирају све вриједности пиксела који представљају објекат ријечи. Дакле, ако је ријеч представљена бијелим пикселима на црној подлози сабира се број бијелих пиксела у једној колони и дијели (нормализује) са висином слике, односно укупним могућим бројем



Слика 3.3: Доњи профил узорка са слике 3.1

бијелих пиксела. Потребно је нагласити да се у случају сивих слика сабирају интензитети свих пиксела у једној колони, при чему се додатно нормализује са максималном вриједношћу сиве слике за одабрани динамички опсег. У сваком случају максимална вриједност овог обиљежја за сваку колону је један. Примјер профила пројекције је дат на Слици 3.4.



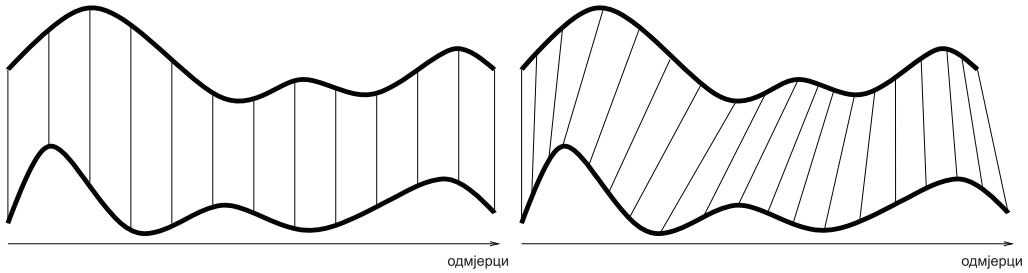
Слика 3.4: Профил пројекције узорка са слике 3.1

3.3 Поређење обиљежја

Профили нису истих дужина због тога што се ријечи екстрахују у апсолутним димензијама, при чему апсолутне димензије слике одређује правоугаоник минималне површине који садржи комплетну ријеч (енг. bounding box). Након екстракције профила (обиљежја), остаје проблем међусобног поређења и оцјене сличности два узорка. Један од начина је да се нормализују дужине профила који су екстраховани, затим да се рачуна, одмјерак по одмјерак, Еуклидска удаљеност. Овакав поступак је осјетљив на локалне деформације које настају приликом писања текста руком, а појављују се у екстрахованим профилима. Управо потреба да се опише сличност узорака са присутним локалним деформацијама је довела до проналаска алгоритма динамичког истезања времена (Dynamic Time Warping - DTW).

3.3.1 Метод динамичког истезања времена - DTW

Разлика између поређења Еуклидском удаљеношћу одмјерак по одмјерак и методом динамичког истезања времена је приказана на Слици 3.5.



(a) Поређење Еуклидском дистанцом (b) Поређење коришћењем DTW-а

Слика 3.5: Различит приступ поређења узорака

Поређење два узорка методом динамичког истезања времена је оптимално у смислу да минимизује кумулативну дистанцу између одмјерака који одговарају упареним тачкама на два узорка. Метод се назива истезање времена због тога што на узорцима деформише (истеже) временске осе тако да се упарени одмјерци позиционирају на исто мјесто заједничке временске осе. Дистанца између два узорка, коришћењем овог метода назива се DTW-дистанца.

Да би се дефинисала DTW-дистанца, посматраће се два узорка $\mathbf{x} = x_1 \dots x_M$ и $\mathbf{y} = y_1 \dots y_N$ која су дужине M и N одмјерака, респективно. DTW-дистанца, означена са $D(M, N)$, се рачуна динамичким програмирањем коришћењем рекурентне једначине

$$D(i,j) = min \left\{ \begin{array}{l} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{array} \right\} + d(x_{i,j}) \tag{3.1}$$

при чему i и j представљају број одмјерка узорака $\mathbf{x}$ и $\mathbf{y}$, респективно, за тренутну итерацију.

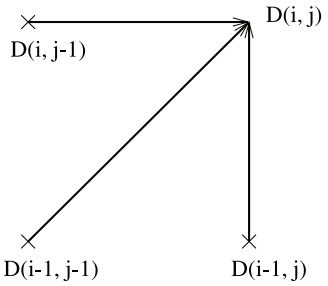
Почетна вриједност DTW-дистанце је једнака нули, а избор функције дистанце унутар парова $d(x_i, y_j)$ је произвољан, бира се зависно од намјене. У овом раду је за $d(x_i, y_j)$ коришћена вриједност квадрата Еуклидске дистанце.

$$d(x_i, y_j) = (x_i - y_j)^2. \tag{3.2}$$

Рекурентна једначина се завршава када је $i = M$ и $j = N$, а представља резултат прорачуна DTW-дистанце.

Избор овакве рекурентне једначине осигурава глатко истезање времена, у смислу да се сваки одмјерак преслика, односно нити један не изостаје приликом истезања. Каже се да оваква рекурентна једначина обезбјеђује услов локалног континуитета, Слика 3.6.

Када се посматра трајекторија парова (i_k, j_k) који имају минималну дистанцу, пролазећи уназад од $D(M, N)$ до $D(1, 1)$ добија се путања истезања (енг. warping path) која има дужину K корака.



Слика 3.6: Илустрација улсова локалног континуитета

3.3.2 Поређење коришћењем DTW-а

Метод динамичког истезања времена у сврху поређења одабраних профила, описаних у претходном поглављу, изводи се сљедећим редослиједом (процедура за један тип профила):

---

1. Израчуна се одабрани тип профила за двије ријечи l, m.
2. Израчуна се међусобна дистанца за та два профила према једначини (3.1) и добије се DTW-дистанца слика l и m, за одабрани тип профила, односно

$$D^{l,m}_{tip\ profila}(M, N) \tag{3.3}$$

   Овај израз се може и скраћено писати на сљедећи начин $D^{l,m}_{tip\ profila}$, при чему се у нотацији изостављају дужине узорака.
3. Нормализује се дистанца (3.3) са дужином путање увијања K и добије се нормализована DTW-дистанца слика l и m, за одабрани тип профила

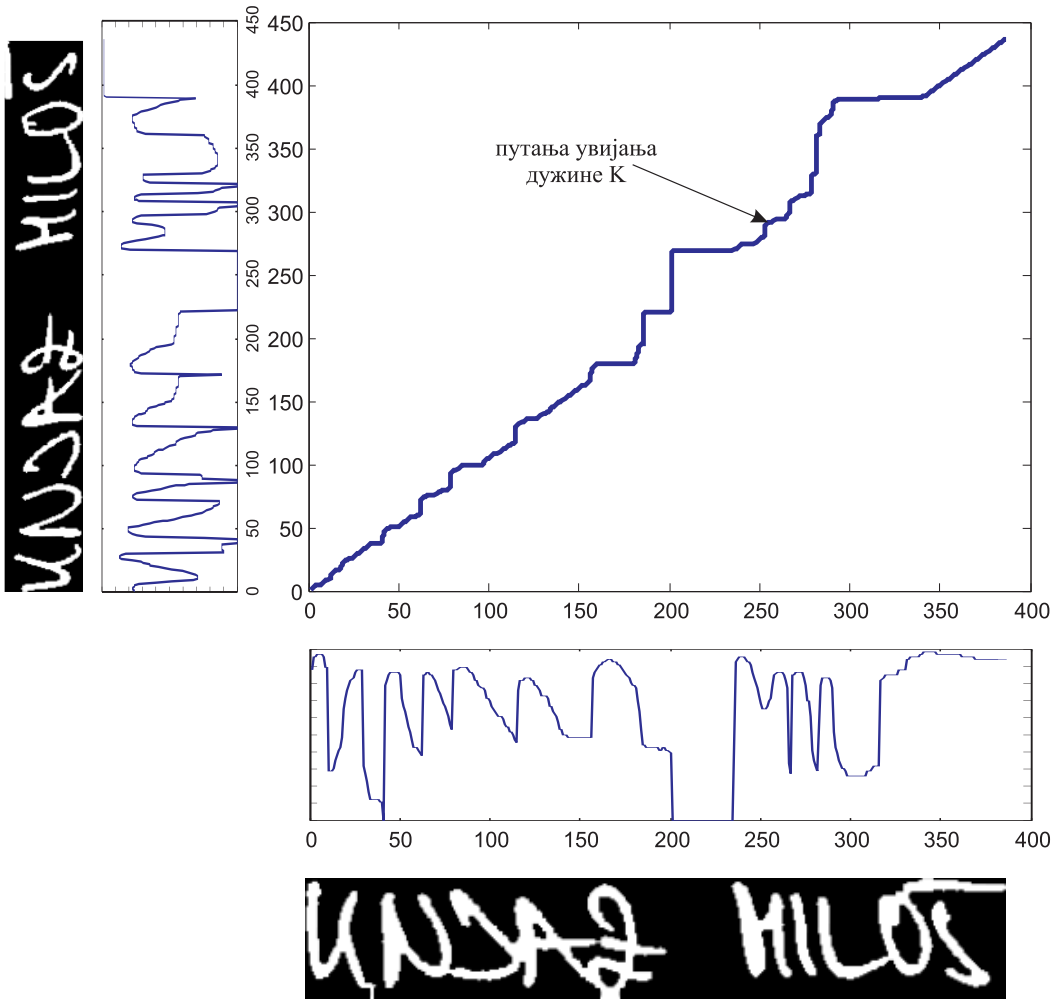
$$D^{l,m}_{N_{tip\ profila}}(M, N) = \frac{1}{K} D^{l,m}_{tip\ profila}(M, N) \tag{3.4}$$

.

Лијева страна израза (3.4) се такође може скраћено писати, изостављајући дужине узорака у самој нотацији, па се добије$D^{l,m}_{N_{tip\ profila}}$.

Пошто је очито да дистанца представљена са (3.3), израчуната из једначине (3.1) зависи од дужине профила, односно од апсолутних димензија слике ријечи, потребно је нормализовати на начин да се поништи наведена зависност. Нормализацијом са дужином путање увијања K поништава се зависност од апсолутних димензија.

На Слици 3.7 је приказан примјер поређења горњег профила за два узорка, коришћењем DTW-а. Индекси ових узорака су 1 и 67, респективно. Један узорак је дужине 387 одмјерака, док је други дужине 436 одмјерака. Дужина путање увијања K за овај случај износи 607 корака. Вриједност ненормализоване дистанце за овај примјер износи 6.5261, па се након нормализације са дужином путање увијања, према једначини (3.4), добије $D^{1,67}_{N_1}(387, 436) = 0.0108$. Са индексом један, као тип профила означен је горњи профил.



Слика 3.7: Поређење горњег профила два узорка коришћењем DTW-а

3.3.3 Матрица дистанци

Када се процедура за поређење коришћењем DTW-а примјени за све могуће парове из скупа узорака добије се матрица дистанци која описује сличност између било која два узорка датог скупа. Вриједности које се уносе у матрицу дистанци су добијене према једначини (3.4), дакле, нормализоване вриједности.

Ова матрица је квадратна, димензија величине скупа узорака. На главној дијагонали ова матрица има вриједности нула, обзиром да је ријеч о дистанци између два идентична узорка. Поред тога, матрица је симетрична у односу на главну дијагоналу, па се често представља само као горња или доња троугаона матрица полазне матрице дистанци. Ради се само о једноставнијем представљању матрице која практично има све елементе позитивне, осим оних на главној дијагонали.

Матрица дистанци је полазна тачка за кластеризацију узорака у функцији израчунатих дистанци, односно класификацију полазног скупа на класе истих ријечи, што је крајњи циљ који се жели постићи.

Потребно је нагласити да за свако обиљежје (горњи профил, доњи профил, профил пројекција) се рачуна једна матрица дистанци. Да би се ови подаци даље обрађивали (кластеризовали) постоје двије могућности. Једна могућност је да се израчунате дистанце у простору сваког обиљежја посматрају као једна димензија вишедимензионог простора па да се у процес кластеризације уђе са вишедимензионалним векторима. Друга могућност је да се редукују подаци на начин да се израчунате дистанце обиљежја комбинују у одређеном смислу и да се добије једна квадратна матрица дистанци која има димензију укупног броја узорака.

У практичном дијелу овог рада коришћена је друга могућност, при чему су матрице дистанци обиљежја комбиноване на начин да се множе елементи матрица дистанци члан по члан за свако обиљежје.

Принципи и алгоритам кластеризације овако добијене матрице су представљени у наредној глави.

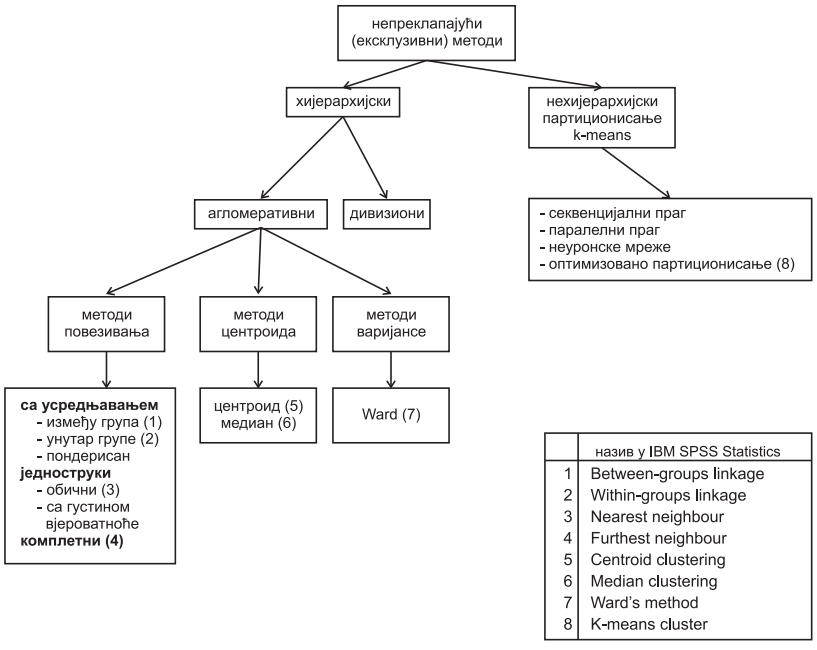
Глава 4

КЛАСТЕРИЗАЦИЈА УЗОРАКА

4.1 Увод

Након одређивања матрице дистанци, потребно је груписати узорке који садрже написану исту ријеч. На Слици 4.1, која је преузета из литературе [12], приказан је преглед метода кластеризације. На истој слици су индексима означене методе које су имплементиране у програмском пакету IBM SPSS Statistics и повезане са претходно поменутим прегледом.

У наставку ће акцент бити стављен на хијерархијске, агломеративне методе. Често цитиран рад аутора Lance-a i Williams-a је систематизовао већину агломеративних метода и дефинисао генерализовану једначину за израчунавање мјере различитости (дистанце) између група кластеризованих узорака [1] .



Слика 4.1: Преглед метода кластеризације, преузето из [12]

За потребе практичног дијела овог рада је коришћена хијерархијска кластеризација и то један од агломеративних метода са усредњавањем између група обиљежја, у литератури познат као метод упаривања непондерисаних група са усредњавањем - Unweighted Pair Group Method with Arithmetic Mean (UPGMA) [1].

На слици 4.1 то је метод са индексом 1, а у програмском пакету SPSS познат је под именом Between-groups Linkage, односно BAVERAGE.

4.2 Агломеративна хијерархијска кластеризација

Код ове групе метода, скупине узорака (кластери) се формирају спајањем узорака, односно спајањем претходно спојених мањих група узорака. Већина ових метода је систематизована радом Lance-a i Williams-a [1]. Поред Lance-a i Williams-a истиче се и рад Murtagh-a [3], који је пронашао постојање временски-оптималног (према броју операција) алгоритма за извршавање наведених метода. Објашњења у овом поглављу се ослањају на ова два рада.

4.2.1 Основне поставке

Посматраће се скуп са N дескрипционих вектора, при чему сваки има M димензија. Дистанца d се може дефинисати на сваком пару ових вектора a и b. У том случају би требало да су задовољене сљедеће три особине:

1. $d(a,b) \geq 0$
2. $d(a,b) = d(b,a)$
3. $d(a,b) \leq d(a,c) + d(b,c)$

Када посљедњи услов (неједнакост троугла) није испуњен, јавља се несличност (dissimilarity, према [3]).

4.2.2 Формула Lance-a i Williams-a за ажурирање несличности

Посматраће се двије групе објеката i и j, са n_i и n_j бројем објеката унутар групе, респективно. Са d_{ij} је означена мјера несличности која се назива међу-групна дистанца. Претпоставиће се даље да је d_{ij} најмања мјера у систему који се посматра, тако да се i и j спајају и чине нову групу k са $n_k = n_i + n_j$ елемената. Посматраће се трећа група h која чини све остале објекте, односно групе објеката, изузимајући објекте из групе k. Формула за ажурирање несличности сада гласи

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}| \tag{4.1}$$

при чему су α_i, α_j, β и γ одређени типом стратегије кластеризације .

Потребно је истакнути да при вриједности $\gamma = 0$ низ дистанци (мјера несличности), израчунатих за свако наредно хијерархијско спајање (агломерацију), ће бити монотон ако је испуњен услов

$$\alpha_i + \alpha_j + \beta \geq 1 \tag{4.2}$$

.

Потребно је нагласити да се у литератури [1], [3] налазе прегледни прикази свих стратегија кластеризације који су систематизовани једначином (4.1), у зависности од

$\alpha_i, \alpha_j, \beta$ и γ. У наставку ће детаљно бити приказан један тип стратегије који се користи у практичном дијелу рада.

4.2.3 Поступак кластеризације

Коришћењем формуле за ажурирање Lance-a i Williams-a, може се поступити на сљедећи начин:

1. Дефинише се прво несличност између парова узорака. Ово може да буде Еуклидска дистанца. У случају практичног дијела овог рада коришћена је нормализована DTW-дистанца.

2. Након прве агломерације два најмање различита објекта, као и код свих наредних агломерација, израчунавају се несличности између кластера и/или преосталих објеката. У ову сврху се користи формула Lance-a i Williams-a, са параметрима специфичним за одређену стратегију (тип) кластеризације.

Потребно је истакнути да ће се због честе заступљености у литератури термини несличност и дистанца користити као идентични појмови, иако формално нису испуњени сви услови да да појам несличности називамо дистанцом.

4.2.4 UPGMA стратегија кластеризације

Ову стратегију кластеризације су први пут формулисали Robert Sokal и Charles Michenery, у раду објављеном 1958. за агломерацију једног елемента и групе елемената. Lance i Williams, 1964. године, уопштавају метод за спајање двије групе елемената [1].

Имајући у виду наведене напомене, прорачун мјера дистанци код UPGMA третираће се као специјалан случај формуле (4.1), при чему су вриједности коефицијената

$$\alpha_i = \frac{n_i}{n_k}; \qquad \alpha_j = \frac{n_j}{n_k}; \qquad \beta = \gamma = 0.$$

Формула за ажурирање међу-групне дистанце (мјере несличности) за овај случај има сљедећи облик

$$d_{hk} = \frac{n_i}{n_k} d_{hi} + \frac{n_j}{n_k} d_{hj} = \frac{n_i \cdot d_{hi} + n_j \cdot d_{hj}}{n_i + n_j} \tag{4.3}$$

.

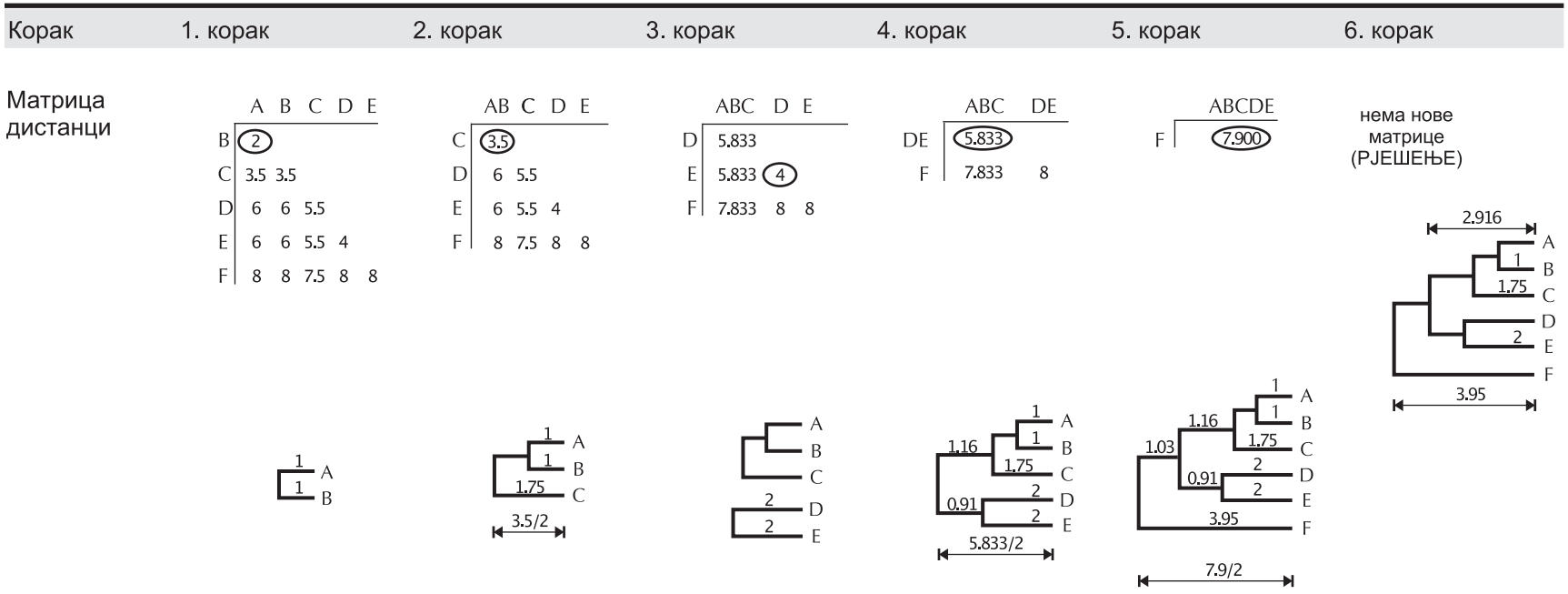
Напоменимо да је чест случај у литератури да се умјестно ознаке k додају претходно спојени индекси којим су означене колоне матрице дистанци. У том смислу код прве агломерације, за $n_i = n_j = 1$, можемо написати $d_{hk} = d_{h(ij)}$. Ово ће бити јасније у практичном примјеру UPGMA кластеризације који слиједи.

4.2.4.1 Илустрација поступка

На Слици 4.2 је илустративно приказан поступак кластеризације UPGMA стратегијом. у наставку слиједи опис приказаних корака, са наведеним прорачуном и формулама.

1. корак

Полази се од матрице дистанци, која је представљена доњом подматрицом ради прегледности. Пронађе се најмања дистанца међу узорцима (пошто још увијек не



Слика 4.2: Илустрација UPGMA стратегије кластеризације

постоји нити једна групу узорака). У овом случају се види да је најмања дистанца 2, између узорака A и B. Потребно је спојити ова два узорка и сажети матрицу дистанци. Формира се дио дендрограма који спаја узорке A и B. Како је овдје ријеч о дистанци између узорака, приликом формирања дијела дендрограма се не користи формула (4.3), већ се директно читају вриједности из матрице дистанци. Поступак спајања A и B дат је у наредном кораку.

2. корак

Да би се добила матрица дистанци приказана у 2. кораку, примјењује се формула (4.3) на сљедећи начин

$$d_{C(AB)} = \frac{n_A \cdot d_{CA} + n_B \cdot d_{CB}}{n_A + n_B} = \frac{1 \cdot 3,5 + 1 \cdot 3,5}{1 + 1} = 3,5$$

аналогно,

$$d_{D(AB)} = \frac{n_A \cdot d_{DA} + n_B \cdot d_{DB}}{n_A + n_B} = 6$$

$$d_{E(AB)} = \frac{n_A \cdot d_{EA} + n_B \cdot d_{EB}}{n_A + n_B} = 6$$

$$d_{F(AB)} = \frac{n_A \cdot d_{FA} + n_B \cdot d_{FB}}{n_A + n_B} = 8$$

Након прорачуна нове матрице дистанци, проналазио минималну дистанцу. Са слике се види да је ријеч о дистанци између новоформиране групе AB и узорка C, а дистанца износи $3,5$. Дендрограм сада изгледа као на Слици 4.2 у 2. кораку. Да бисмо добили матрицу дистанци за сљедећи корак, потребно је израчунати удаљености новонастале групе (кластера) ABC према свим осталим елементима, односно групама.

3. корак

Да би се добила матрица дистанци приказана у 3. кораку, примјењује се формула (4.3) на сљедећи начин

$$d_{D(ABC)} = \frac{n_{AB} \cdot d_{D(AB)} + n_C \cdot d_{DC}}{n_{AB} + n_C} = \frac{2 \cdot 6 + 1 \cdot 5,5}{2 + 1} = 5,833$$

аналогно,

$$d_{E(ABC)} = \frac{n_{AB} \cdot d_{E(AB)} + n_C \cdot d_{EC}}{n_{AB} + n_C} = \frac{2 \cdot 6 + 1 \cdot 5,5}{2 + 1} = 5,833$$

$$d_{F(ABC)} = \frac{n_{AB} \cdot d_{F(AB)} + n_C \cdot d_{FC}}{n_{AB} + n_C} = \frac{2 \cdot 8 + 1 \cdot 7,5}{2 + 1} = 7,833$$

а остали елементи матрице дистанци, која је приказана у 3. кораку, се преписује из претходне. Сада се види да је минимална дистанца у овој матрици она између узорака D и E, а има вриједност 4. Дакле у овом кораку се спајају узорци D и E у нови кластер, а дендрограм изгледа као на Слици 4.2 у 3. кораку.

4. корак

Да би се добила матрица дистанци приказана у 4. кораку, потребно је израчунати дистанцу новонасталог кластера DE према осталим кластерима и узорцима (у

овом случају према кластеру ABC и узорку F. примјењује се формула (4.3) на сљедећи начин

$$d_{(ABC)(DE)} = \frac{n_D d_{(ABC)D} + n_E d_{(ABC)E}}{n_D + n_E} = \frac{1 \cdot 5,833 + 1 \cdot 5,833}{1+1} = 5,833$$

$$d_{F(DE)} = \frac{n_D d_{FD} + n_E d_{FE}}{n_D + n_E} = \frac{1 \cdot 8 + 1 \cdot 8}{1+1} = 8$$

Дистанца $d_{F(ABC)}$ се преписује из претходног корака и добије се жељена матрица дистанци. Видимо да је минимална дистанца $5,833$ између кластера ABC и DE, па су на дендрограму четвртог корака управо ови кластери спојени.

5. корак

Да би се добила матрица дистанци приказана у 5. кораку, остаје да се израчуна удаљеност новог кластера $ABCDE$ од узорка F. Примјењује се формула (4.3) на сљедећи начин

$$d_{F(ABCDE)} = \frac{n_{ABC} d_{F(ABC)} + n_{DE} d_{F(DE)}}{n_{ABC} + n_{DE}} = \frac{3 \cdot 7.833 + 2 \cdot 8}{3+2} = 7,90$$

Након спајања кластера $ABCDE$ и узорка F добија се дендрограм као на слици у петом кораку.

6. корак

Коначно, види се да нема нове матрице дистанци која се може креирати и каже се да је са овим процес завршен. Дендрограм са дистанцама тачака агломерације је приказан као коначно рјешење. Детаљне дистанце сегмената дендрограма се виде у петом кораку, пошто је ријеч о идентичном дендрограму.

4.2.4.2 UPGMA у програмском пакету SPSS

Да би се илустровало коришћење UPGMA стратегије у програмском пакету IBM SPSS Statistics (SPSS), користиће се идентична матрица дистанци као из претходне илустрације.

Дакле, улазна матрица је идентична оној која је приказана у првом кораку на Слици 4.2, а дефинисана је у датотеци `MDilustr.sav`.

```
CLUSTER

  /MATRIX IN  ('C:\MDilustr.sav')
  /METHOD BAVERAGE

  /PRINT SCHEDULE
  /PLOT DENDROGRAM.
```

Слика 4.3: Програмска секвенца за UPGMA кластеризацију у SPSS-у

На Слици 4.3 приказана је програмска секвенца (скрипта) којом се у SPSS-у кластеризује матрица дистанци коришћењем UPGMA стратегије. У наставку ће се нешто

детаљније описати значење садржаја ове програмске секвенце и детаљно ће бити образложени резултати које SPSS генерише.

Прва линија ове програмске секвенце `CLUSTER` означава команду да ће се извршавати кластеризација података.

Наредна линија секвенце одређује карактер улазних података, односно дефинише да је улаз матрица дистанци, а не подаци међу којима је потребно прво одредити дистанце. Поред тога, овај дио програмске секвенце дефинише над којом матрицом ће се извршавати кластеризација, односно дефинише путању до датотеке која садржи матрицу дистанци.

Трећа линија програмске секвенце дефинише метод кластеризације. Када се постави параметар BAVERAGE, ријеч је о UPGMA стратегији кластеризације, коју SPSS назива Between-groups Linkage, односно Average Linkage Between Groups.

Наредне двије линије служе за испис резултата. Прва исписује распоред агломерације (Agglomeration Schedule), док посљедња линија секвенце исцртава дендрограм рјешења.

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	2	2.000	0	0	2
2	1	3	3.500	1	0	4
3	4	5	4.000	0	0	4
4	1	4	5.833	2	3	5
5	1	6	7.900	4	0	0

Слика 4.4: Агломерациони распоред рјешења

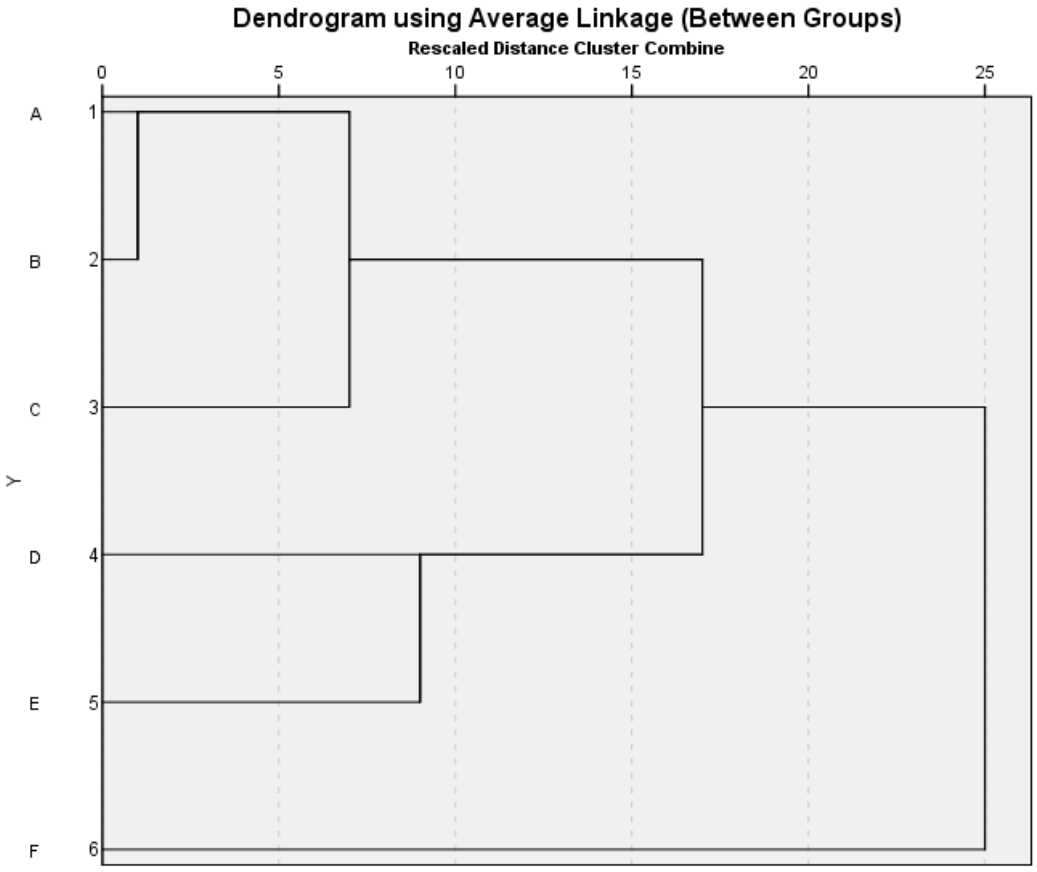
Агломерациони распоред је табеларни приказ карактеристичних података, везаних за поступак кластеризације, које генерише SPSS. На Слици 4.4 је приказан агломерациони распоред за матрицу дистанци приказаних у поступку са Слике 4.2.

Агломерациони распоред приказује неколико занимљивих података. У првој колони се налази редни број корака поступка (Stage). Четврта колона (Coefficients) приказује вриједност минималне дистанце за дати корак. Вриједности из ове колоне се подударају са оним добијеним у претходно описаном поступку. Друга и трећа колона представљају кластере који се спајају на датом кораку и то тако да новонастали кластер узима вриједност коју је прије спајања имала вриједност друге колоне (Cluster 1). Посљедња колона (Next Stage) приказује редни број корака у којем спојени кластери, у датој врсти, поново учествују у спајању. Пета и шеста колона приказују када су се наведени кластери који се спајају први пут појавили, односно редни број корака када су они први пут настали.

Увидом у агломерациони распоред за наведену матрицу дистанци, можемо потврдити да је ријеч о UPGMA поступку, на основу вриједности дистанци за сваки корак агломерације које смо добили рачуном претходном поглављу. Агломерациони распоред

нуди многе занимљиве податке на прегледан начин, ипак не показује о којим узорцима и групама је ријеч. Увид у ове информације на веома прегледан начин нам показује дендрограм.

Прегледан приказ груписања узорака и кластера у процесу кластеризације се види на дендрограму. Дендрограм приказан на Слици 4.5 је генерисан са претходно приказаном програмском секвенцом.



Слика 4.5: Дендрограм рјешења

Овај дендрограм тополошки одговара оном који је приказан као резултат у поступку на Слици 4.2, али види се да вриједности дистанци не одговарају по апсолутним износима. Разлог лежи у начину на који SPSS приказује дендрограм. Након прорачуна вриједности дистанци као што је учињено на Слици 4.2, SPSS линеарно скалира добијене вриједности дистанци у опсег од 1 до 25. Други начин на који можемо гледати ово пресликавање је да посматрамо распоред агломерације, приказан на Слици 4.4, и то колону Coefficients. Скалирање које SPSS извршава линеарно пресликава вриједност једне половине првог елемента те колоне у вриједност 1, а вриједност једне половине посљедњег елемента у вриједност 25.

Након наведеног скалирања, SPSS на крају квантује вриједности дистанци на цијеле бројеве од 1 до 25 и на тај начин приказује дендрограм.

Управо здружени увид у распоред агломерације и дендрограм даје праву слику

о процесу кластеризације. Из дендрограма се види тополошки резултат, а из агломерационог распореда егзактне вриједности дистанци добијених у корацима UPGMA поступка.

Потребно је нагласити да су, поред ових, доступни многи додатни подаци везани за процес кластеризације који се приказују одређеним командама, али у овом примјеру је програмска секвенца написана минималистички да би се истакла главна својства процеса и презентовао приказ дендрограма у SPSS-у.

Глава 5

ПРАКТИЧНА РЕАЛИЗАЦИЈА И ПРЕГЛЕД РЕЗУЛТАТА

За практичне потребе рада узети су анкетни листови које је испунио један анкетар у истраживању спроведеном од стране фирме D2D Advertising. Анкетни листови су скенирани као црно-бијеле слике резолуцијом 300 тачака по инчу. Предобрада ових докумената је реализована као дио комерцијалног софтвера који је аутор овог рада развијао претходних неколико година. Ова предобрада укључује геометријске као и одређене морфолошке операције.

Са тако обрађених анкетних листова узет је скуп од 112 узорака, односно изоловане су слике ријечи које је, као што је напоменуто, написао један анкетар. Овако добијене слике имају различите просторне димензије.

Наставак обраде ових слика, што је предмет овог рада, је реализован у програмском пакету Matlab. Кластеризација и презентација резултата су реализовани у програмском пакету IBM SPSS Statistics (SPSS). Програмске секвенце су дате као прилог на CD-у који прати овај рад.

За сваку слику су израчунати горњи профил, доњи профил, као и профил пројекција. Након поређења коришћењем нормализоване DTW, конструисане су матрице дистанци димензија 112×112 за сваки од израчунатих профила. Ове три матрице су међусобно помножене члан по члан, па се на тај начин добила матрица дистанци димензија 112×112, реалних коефицијената, над којом се даље вршила кластеризација.

Презентација резултата процеса кластеризације је дата у виду дендрограма. Дендрограм овог процеса је приказан у Прилогу А, при чему је поред назива и редног броја узорка приказана и његова слика.

Потребно је напоменути да 11 узорака није добро сегментирано, односно да је на ивицама слика на неким мјестима остао објекат, који је тополошки измјенио садржај слике. Интересанто је да је поступак кластеризације све ове случајеве спојио у један кластер на нормализованој дистанци SPSS дендрограма која има вриједност 10. Ово се објашњава чињеницом да су практично сви узорци имали погрешно сегментиране аномалије у доњем десном дијелу слике, па су израчунати профили, тополошки посматрано, били веома слични.

Што се тиче грешака кластеризације, види се да су практично само четири узорка погрешно кластеризовани у смислу да су неприродно спојени или да не припадају одређеном кластеру.

За крај, посматраће се исправно кластеризовани узорци. Као што се види, велики број узорака је исправно кластеризован.

Оно што је од највећег практичног значаја јесте чињеница да се за овај скуп узорака, са шест радних операција може тачно означити шест група узорака које имају исти садржај унутар групе (укупно 73 узорка). Ако се посматра дендрограм у Прилогу А, види се да су групе ријечи: Кисељак (10), Кназ М(19), Врњци (7), Раденска (6), Витинка (15) и Кназ Милош (16) тополошки груписане тако да из сваке од наведених група само једна грана повезује све узорке групе са наредним хијерархијским нивоом. Другим ријечима, ако би се дало оператеру да одабере (кликне на) ту грану, аутоматски би се исправно означили сви узорци који припадају одговарајућој групи.

Дакле, ако искључимо грешке сегментације, приказом дендрограма могуће је са шест радних операција тачно означити нешто више од 72 одсто укупног броја узорака. Комплетан скуп би се тачно означио са 23 радне операције укључујући разрјешење свих грешака, што је готово петина (22,7 одсто) од укупног броја радних операција које су потребне за директан унос (не рачунајући грешке сегментације). Потребно је напоменути да би се поступак морао поновити за сваког анкетара, што не представља велику деградацију претходно поменутих резултата обзиром на релативно велик број анкета по једном анкетару.

Поред могућности много бржег уноса, јасно је да се видно смањује грешка уноса пошто су груписани узорци са истим текстом, тако да би се погрешан узорак јако истицао.

Глава 6

ЗАКЉУЧАК

У раду је приказан поступак који омогућава полуаутоматизован унос текстуалних података који су на папиру писани руком а припадају затвореном скупу података. Приказан поступак претпоставља да је све ријечи руком написала једна особа - анкетар.

Поступак има највећу примјену и вриједност у случају да се уносе подаци који чине релативно мали скуп различитих вриједности, при чему се те вриједности бирају из много већег скупа.

Такав случај даје велику предност поступку приказаном у овом раду, у односу на директан (мануелни) унос података. Поред тога, природа функционисања и презентације (дендрограм) омогућава унос са много мањом грешком.

Са резултатима приказаним у раду се види да је брзина уноса оваквих података повећана за готово пет пута, при чему је тачност унесених података повећана више од четири пута у односу на мануелни поступак (на много прегледнији начин је потребно унијети 24 одсто података, што практично значи мању вјероватноћу грешке).

Постоји много могућности за развој и усавршавање приказаног поступка. Прије свега потребно је квалитетније ријешити предобраду у смислу да се идентификују и ријеше грешке сегментације. Ово није урађено у практичном дијелу рада, али са данашњим сазнањима представља реално лако рјешив проблем.

Поред предобраде потребно је истражити додатна обиљежја и анализирати могућност усавршавања екстракције постојећих. Као примјер, наводи се методологија екстракције обиљежја коришћењем Harris-овог детектора ћошкова [10].

Кластеризацију и презентацију је могуће прилагодити кориснику на начин да након идентификације неколико кључних кластера, са великим бројем узорака, поновимо процедуру за преостале узорке са могућим коришћењем одређених поступака машинског учења. У том случају систем већ посједује информације о класификацији за које „тврдимо" да су истините, па је већа вјероватноћа исправне класификације преосталих узорака.

ЛИТЕРАТУРА

[1] G. N. Lance and W. T. Williams. A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems. *The Computer Journal*, 9(4):373–380, feb 1967.

[2] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, November 2004.

[3] F. Murtagh. Complexities of hierarchic clustering algorithms: State of the art. *Computational Statistics Quarterly*, 1:101–113, 1984.

[4] J. J. D. Oliveira, Jr., J. M. de Carvalho, C. O. D. A. Freitas, and R. Sabourin. Feature sets evaluation for handwritten word recognition. In *Proceedings of the Eighth International Workshop on Frontiers in Handwriting Recognition (IWFHR'02)*, IWFHR '02, pages 446–, Washington, DC, USA, 2002. IEEE Computer Society.

[5] R. Plamondon and S. N. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:63–84, January 2000.

[6] K. Pramod Sankar, C. V. Jawahar, and R. Manmatha. Nearest neighbor based collection OCR. In *DAS '10: Proceedings of the 8th IAPR International Workshop on Document Analysis Systems*, pages 207–214, New York, NY, USA, 2010. ACM.

[7] T. M. Rath and R. Manmatha. Features for word spotting in historical manuscripts. In *Features for Word Spotting in Historical Manuscripts*, volume 1, page 218, Los Alamitos, CA, USA, 2003. IEEE Computer Society.

[8] T. M. Rath and R. Manmatha. Word image matching using dynamic time warping. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:521, 2003.

[9] J. A. Rodriguez and F. Perronnin. Local gradient histogram features for word spotting in unconstrained handwritten documents. In *Proceedings of the 1st International Conference on Handwriting Recognition (ICFHR'08)*, aug 2008.

[10] J. L. Rothfeder, S. Feng, and T. M. Rath. Using corner feature correspondences to rank word images by similarity. *Computer Vision and Pattern Recognition Workshop*, 3:30, 2003.

[11] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49, 1978.

[12] M. Schmidt and S. Hollensen. *Marketing Research: An International Approach*. Prentice Hall, 2006.

ПРИЛОГ А

ГРАФИЧКИ ПРИКАЗ РЕЗУЛТАТА