yhden pikselin päässä oleva vastine toisessa joukossa. Vain jos näin ei ollut, laskettiin pisteen etäisyys kaikkiin toisen joukon alkioihin. Laskenta-aika lyheni tällä ratkaisulla yli 70%. Muita etäisyysmittojen laskennan optimointimenetelmiä käsitellään mittojen kuvauksissa.

3.2 Vektorietäisyys

Tässä tutkielmassa käsitellään ihmisen luettavaksi tarkoitettuja kirjoitusmerkkejä, kuten länsimaisia numeroita ja kirjaimia. Niiden erottamisessa toisistaan ei piirretyn viivan harmaasävyllä ole merkitystä, vaan merkki koostuu pikseleistä, jotka joko ovat päällä tai eivät ole. Kuva on mustavalkoinen. Jos kuvaa tarkastellaan vektorina, kutakin kuvan pikseliä vastaa yksi vektorin alkio, joka saa siten jommankumman mahdollisista arvoista, esim. 1 tai 0. Kahden tällaisen vektorin euklidinen etäisyys on arvoja 1 ja 0 käyttäen yhtä kuin vektorien toisistaan poikkeavien alkioiden lukumäärän neliöjuuri. Tässä tutkielmassa nimitetään *vektorietäisyydeksi* näiden alkioiden lukumäärää sellaisenaan, koska se tuottaa saman etäisyysjärjestyksen havaintoon verrattavien mallien välille. Mitta on myös metriikka. Ei ole mahdollista, että hahmot A ja B eroaisivat toisistaan jonkin pikselin (i, j) osalta, mutta kolmannen hahmon C pikseli (i, j) olisi sama sekä A:n että B:n vastaavan pikselin kanssa. Näin etäisyys ei voi pienetä kuljettaessa kolmannen objektin kautta. Eroavien pisteiden määrä on triviaalisti epänegatiivinen ja se on nolla, jos ja vain jos hahmot ovat samat. Laskentakaava on siten seuraava:

$$d_v(A, B) = |A \setminus B| + |B \setminus A| \quad .$$

Kun vektoria ajatellaan pistejoukon kuvauksena, on vektorimitta pistejoukkojen etäisyysmitta. Muista käsiteltävistä mitoista poiketen vektorietäisyys on mielekäs vain äärellisen ja suppean diskreetin avaruuden pistejoukoille. Muuten kaksi joukkoa voisivat olla vektorimitan mukaan hyvin kaukana toisistaan, vaikka niillä olisi sama määrä pisteitä, ja hyvin lähellä jokaisen ensimmäisen joukon pistettä olisi jokin toisen joukon piste. Vektorietäisyys on laskettavissa hyvin nopeasti, kun kuvat talletetaan sopivasti, sillä vastinpisteiden vertaaminen riittää etäisyyden laskemiseksi.

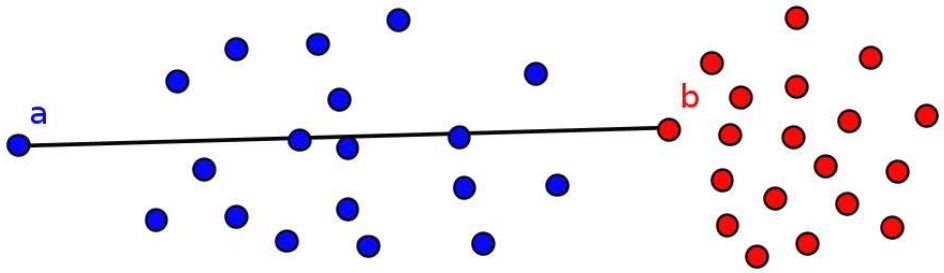
3.3 Hausdorff-etäisyys

Saksalaisen matemaatikon Felix Hausdorffin mukaan nimetty *Hausdorff-etäisyys* kertoo kuinka kaukana enimmillään on jokin jommankumman joukon alkio sitä lähimmästä toisen joukon alkiosta [Hau14]. Hausdorff esitti etäisyysmitan kaikille

metrisen avaruuden osajoukoille, jolloin laskentakaavassa käytettiin tässä esiintyvän minimin sijasta infinumia. Laskentakaava kaikille kompakteille joukoille ja siten erityisesti tässä tutkielmassa käsiteltäville äärellisille joukoille on

$$d_h(A,B) = \max\left(\max_{a\in A} \min_{b\in B} d(a,b), \max_{b\in B} \min_{a\in A} d(b,a) \right) .$$

Hausdorff-etäisyys eli *Hausdorff-metriikka* on mitta, jota voidaan soveltaa, kun halutaan nähdä myös kaksi aivan eri kokoista pistejoukkoa samankaltaisena, kunhan kummankin joukon kukin alkio on lähellä jotain toisen joukon alkiota. Se ei yleisesti kuvaa lainkaan joukkojen kokoa eikä niiden rakenteen tai koon samankaltaisuutta. Kuvassa 3.1 on esimerkki tilanteesta, jossa kaksi pistettä määrää joukkojen välisen etäisyyden. Kaikki muut pisteet voisi poistaa etäisyyden muuttumatta.



Kuva 3.1 Tilanne, jossa pisteet a ja b määräävät yksin joukkojen välisen Hausdorff-etäisyyden.

Hausdorff-etäisyys on suuri, jos yksikin piste on kaukana kaikista toisen joukon pisteistä, vaikka joukot olisivat identtiset tätä yhtä alkiota lukuun ottamatta. Jotta luokittelu onnistuu, ei havaintoaineistossa saa siten olla lainkaan mittausvirheitä, jotka ovat merkittäviä suhteessa havainnon ja mallin välisiin todellisiin eroihin. Vastaavasti tarvitaan jokaista tiettyyn luokkaan kuuluvaa havaintoa kohti sellainen tähän luokkaan kuuluva malli, jossa ei ole yhtään isoa poikkeamaa havaintoon verrattuna. Mikään muu pistejoukkojen etäisyysmitta ei ole yhtä herkkä yksittäiselle virheelle tai poikkeamalle vertailtavien joukkojen välillä. Hausdorff-etäisyys on usein käyttökelpoinen mitta luokittelussa ja muissa tehtävissä, joissa tavoitteena on melko läheisten joukkojen tunnistaminen ja samankaltaisuuden määritelmä on edellä kuvatun kaltainen. Esimerkiksi malliin perustuva kuvantunnistus valokuvasta on tällainen tehtävä [HKR93]. Toisistaan etäisiä pistejoukkoja tämä etäisyysmitta ei aseta intuitiiviseen erilaisuusjärjestykseen.

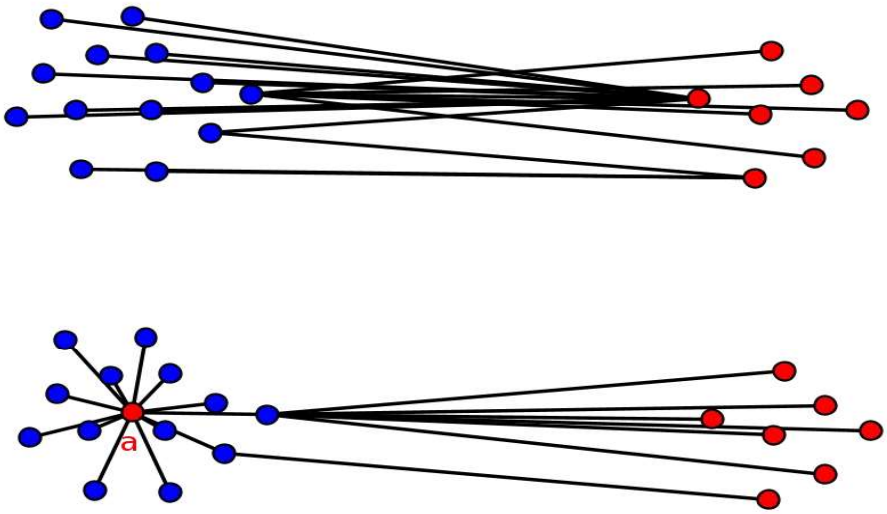
Useimmista muista etäisyysmitoista poiketen Hausdorff-etäisyys on metriikka [Hau14]. Täten jos se soveltuu tutkittavan ilmiön samankaltaisuuden määritelmäksi, voidaan luokittelussa käyttää mallien metristä indeksointia vähentämään olennaisesti tarvittavien vertailujen määrää. Hausdorff-etäisyyden suoraviivainen laskenta on aikavaativuudeltaan O(m*n), mutta sen laskemiseksi ei tarvitse selvittää jokaisen alkion etäisyyttä jokaiseen toisen joukon alkioon kuin pahimmassa tapauksessa. Kun havaitaan, että alkion etäisyys toiseen joukkoon on pienempi tai yhtä suuri kuin jonkin muun alkion etäisyys toiseen joukkoon, voidaan vertailu tämän alkion osalta lopettaa. Lisäksi voidaan käyttää optimointitapoja, jotka soveltuvat kaikille lähimmän vastinalkion etäisyyteen perustuville mitoille. Kuvantunnistuksen kannalta on erittäin hyödyllistä, jos etäisyysmitalle voidaan tehokkaasti löytää pienin mahdollinen arvo, joka saavutetaan siirtämällä toista joukkoa avaruudessa. Pienin Hausdorff-etäisyys yhdensuuntaissiirrossa on laskettavissa kaksiulotteisessa tapauksessa ajassa $O(n^4)$ [HuK90].

3.4 Minimietäisyyksien summa

Minimietäisyyksien summa -mitta lasketaan kaavalla

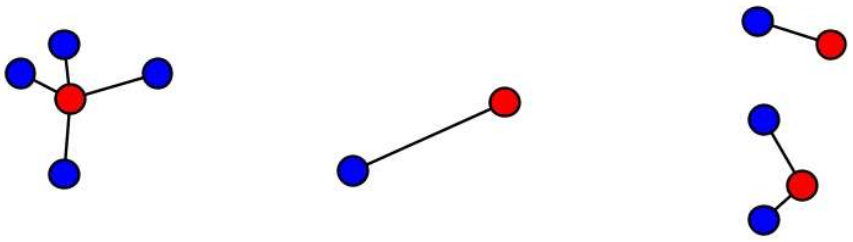
$$d_s(A,B) = \frac{1}{2}\left(\sum_{a\in A} \min_{b\in B} d(a,b) + \sum_{b\in B} \min_{a\in A} d(b,a)\right),$$

eli mitan nimestä huolimatta sen varsinainen arvo on suunnattujen summien keskiarvo [Nii87]. Tällä ei ole vaikutusta siihen, millaiseen etäisyysjärjestykseen minimietäisyyksien summa joukot asettaa. Minimietäisyyksien summan perusteella havaintoa lähellä on malli, jossa kullekin havainnon pisteelle löytyy lähellä oleva pari, mutta ei kovin monia vastinalkioita. Aiemmin kuvattuun Hausdorff-etäisyyteen ei vastinpisteiden määrä vaikuta. Pisteen lisääminen joukkoon A pienentää joukkojen A ja B välistä etäisyyttä tai säilyttää sen, jos uudesta pisteestä tulee yhden tai useamman B:n pisteen lähin piste joukossa A, ja näiden pisteiden etäisyys joukkoon A pienenee yhteensä vähintään saman verran kuin on lisätyn pisteen etäisyys joukkoon B. Joukkojen välinen etäisyys voi pienetä olennaisesti, jos moni piste saa paljon läheisemmän parin toiseen joukkoon, kuten kuvassa 3.2. Tällöin joukot ovat melko erilaiset, joten mitta ei silti ole käytännössä herkkä kohinalle luokittelua ajatellen. Muuten joukkojen välinen etäisyys useimmiten kasvaa, jos joukkoihin lisätään alkioita, joiden etäisyys toiseen joukkoon ei ole nolla. Minimietäisyyksien summa kuvaa siis merkittävästi paitsi joukkojen samankaltaisuutta, myös niiden kokoa.



Kuva 3.2: Pisteen a lisääminen punaiseen joukkoon pienentää tässä huomattavasti minimietäisyyksien summaa.

Aivan lähelle toista joukkoa voi aina lisätä suurenkin määrän pisteitä etäisyyden juuri kasvamatta. Jotkin muut etäisyysmitat voivat kasvaa rajustikin, kun pisteitä lisätään näin, ja jopa vaikka uusille pisteille olisi identtinen pari toisessa joukossa. Minietäisyyksien summa ei niiden tapaan kerro, kuinka tasapuolisesti joukon alkiot sijoittuvat toisen joukon eri alkioiden lähettyville. Kuvassa 3.3 on esitetty kolme tapausta, joissa punaisen ja sinisen joukon välinen etäisyys on sama.

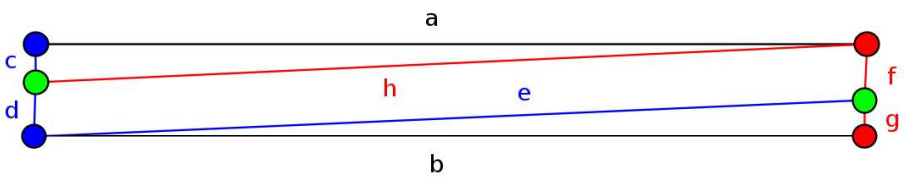


Kuva 3.3: Kolme tilannetta, joissa sinisen ja punaisen joukon etäisyys on minietäisyyksien summan perusteella sama

Jos merkkien tunnistuksessa halutaan nähdä ohut ja paksu viiva selvästi eri ilmiönä, on minimietäisyyksien summa toimiva etäisyysmitta. Pieniin yksityiskohtiin se ei reagoi kovin herkästi, sillä osajoukon pisteiden etäisyys vastakkaiseen joukkoon vaikuttaa etäisyysmitan arvoon suhteessa sitä vähemmän, mitä enemmän muita alkioita joukoissa on. Lisäksi toisen joukon alkioita ei kohdenneta laskennassa ensimmäisen joukon etäisen osajoukon alkioihin, vaan tällainen osajoukko kasvattaa vain toisen suuntaista summaa. Kohinan siedon kannalta pienten etäistenkin pistejoukkojen vähäinen vaikutus on hyvä asia.

Optisessa merkkien tunnistuksessa mitta ei välttämättä tuota parasta mahdollista tulosta ainakaan länsimaisten kirjainten ja numeroiden osalta. Niiden erottelu ei perustu viivan paksuuden vertailuun, mutta melko pienet erot sen sijaan erottavat joitakin kirjaimia toisistaan. Minimietäisyyksien summan mukaan ohuella viivalla piirretty A-kirjain muistuttaa enemmän mallia, jossa on ohuella viivalla piirretty Ä, kuin mallia, jossa on paksulla viivalla piirretty A. Esimerkiksi Hausdorff-etäisyys antaa päinvastaisen tuloksen, kunhan Ä-kirjaimen pisteet ovat riittävän etäällä.

Jotta tunnistus onnistuu, tarvitaan oma malli eri paksuisella jäljellä kirjoitettuja merkkejä varten. Tällöin vertailuja joudutaan tekemään paljon, sillä minimietäisyyksien summa ei ole metriikka [Gär88], joten ei ole helppoa karsia tehokkaasti pois sellaisia malleja, jotka ilman muuta ovat kaukana havainnosta. Kuvassa 3.4 on esimerkki tilanteesta, jossa metriikan vaatimukset eivät täyty. Sinisen ja punaisen joukon välinen suora etäisyys on pidempi kuin sinisen joukon etäisyys vihreästä joukosta ja vihreän joukon etäisyys punaisesta joukosta yhteensä. Janojen a, b, c ja g pituudet huomioidaan laskennassa kaksi kertaa, muiden pituus vain kerran.



Kuva 3.4: Minimietäisyyksien summa ei ole metriikka. Suora etäisyys sinisestä joukosta punaiseen joukkoon, joka on $|a| + |b|$, on pidempi kuin etäisyys vihreän joukon kautta, eli $|c| + |g| + (|d| + |f| + |h| + |e|) / 2$.

Minimietäisyyksien summan laskenta triviaalisti vie ajan O(mn). Jokaisen alkion etäisyys kuhunkin toisen joukon alkioon selviää siinä ajassa, mutta laskentaa voidaan helposti optimoida tekniikoilla, joilla löydetään useimmille alkioille lähin alkio vastakkaisessa joukossa käymättä läpi kaikkia vastakkaisen joukon alkioita.

3.5 Surjektioetäisyys

Surjektioetäisyys [Odd79] kertoo, kuinka helposti isompi joukko voidaan muuntaa pienemmäksi joukoksi, eli miten suuri vastinalkioiden yhteisetäisyys vähintään on joukkojen välisessä surjektiivisessa kuvauksessa. Se siis antaa minimin matkalle, joka kertyy, kun isomman joukon jokainen alkio siirretään jonkin pienemmän joukon alkion kohdalle siten, että kunkin pienemmän joukon alkion kohdalle tuodaan vähintään yksi alkio. Surjektioetäisyys lasketaan seuraavasti:

$$d_{sur}(A,B) = \min_{\eta \in N} \sum_{(a,b)\in\eta} d(a,b) \quad ,$$

missä $|A| \geq |B|$, $a \in A$, $b \in B$ ja N on mahdollisten surjektioiden joukko. Ainakin pelkistetyissä tilanteissa surjektioetäisyys antaa usein vertailukohteille saman järjestyksen kuin minietäisyyksien summa. Etäisyys on kuitenkin merkittävästi erilainen kuvan 3.5 kaltaisessa tilanteessa. Tällöin joukkojen yhdiste sisältää erillisiä pistesaarekkeita, joissa kussakin on molempien joukkojen alkioita, ja joista osassa on pienemmän joukon alkioita enemmän kuin suuremman. Saarekkeiden välinen etäisyys vaikuttaa silloin joukkojen surjektioetäisyyteen, mutta ei minimietäisyyksien summaan, joka voi myös olla paljon pienempi. Tällainen tilanne on sitä epätodennäköisempi, mitä isompi on joukkojen kokoero. Myös näin syntyvien etäisten vastinparien merkitys kokonaisetäisyyden kannalta vähenee kokoeron kasvaessa.