Kasvojentunnistukseen kohdistuvat imitaatiohyökkäykset ja niiden estäminen

Turun yliopisto
Tietotekniikan laitos
LuK-tutkielma
Tietojenkäsittelytiede
Toukokuu 2026
Otto Viljanen

TURUN YLIOPISTO
Tietotekniikan laitos

Otto Viljanen: Kasvojentunnistukseen kohdistuvat imitaatiohyökkäykset ja niiden estäminen

LuK-tutkielma, 28 s.
Tietojenkäsittelytiede
Toukokuu 2026

Tässä työssä tarkastellaan kasvojentunnistusohjelmien sabotointia kohdistetuilla imitaatiohyökkäyksillä ja niiltä suojautumista. Kohdistetussa imitaatiohyökkäyksessä hyökkääjä pyrkii sabotoimaan kasvojentunnistusohjelman toimintaa ja saaden sen tunnistamaan hyökkääjän imitoinnin kohteena olevaksi henkilöksi. Tutkielmaa varten kerättiin erilaisia imitaatiohyökkäyksiä käsittelevistä artikkeleista aineisto, jonka perusteella osoitettaan kohdistettujen imitaatiohyökkäysten olevan mahdollisia toteuttaa kaikkiin kasvojentunnistusprosessin vaiheisiin, jotka ovat imitaation kannalta merkittäviä. Tämän lisäksi tarkastellaan erilaisia imitaatiohyökkäysten tunnistamisen ja estämisen keinoja. Aineistossa ja muissa lähteissä esitettyjä suojauskeinoja arvioidaan niiden sisältämien heikkouksien, kustannusten ja käyttöönoton sujuvuuden perusteella. Merkittävänä havaintona esille nousi havainto usean suojauskeinon pystyvän tehokkaasti tunnistamaan vain yhden tyypin hyökkäyksiä. Erilaisten imitaatiohyökkäysten läpikäynnin sekä suojauskeinojen arvioinnin aikana esille tuotujen havaintojen pohjalta todetaan fuusiomallien olevan tehokas suojauskeino imitaatiohyökkäyksiltä. Fuusiomallit yhdistävät useita imitaatiohyökkäyksen tunnistamisen ja estämisen metodeja ja täten kattavat suuren joukon erilaisia hyökkäyksiä, mutta niiden kehittäminen ja käyttöönotto vaativat aikaa. Toisena merkittävänä suojauskeinona esille nousee ihmistarkastaja, joka on nopeasti käyttöönotettavana ja samalla tehokas suojauskeino, jos käsiteltävän datan määrä ei ole suuri.

Asiasanat: Kasvojentunnistus, imitaatiohyökkäys, koneoppiminen, suojauskeinot

UNIVERSITY OF TURKU
Department of Computing

Otto Viljanen: Kasvojentunnistukseen kohdistuvat imitaatiohyökkäykset ja niiden estäminen

Bachelor's Thesis, 28 p.
Computer Science
May 2026

This work examines sabotaging face recognition systems with targeted impersonation attacks and defensive techniques against them. A targeted impersonation attack aims to sabotage a face recognition system in a manner that makes the attacker appear as the target of impersonation to the system. Articles covering different impersonation attacks were gathered as material for the thesis which showed that impersonation attack can be aimed at any part of the recognition process that are relevant for impersonation attacks. In addition different methods for identifying and preventing impersonation attacks were examined. Defense methods presented in the material and in other sources are evaluated based on their weaknesses, cost and ease of implementation. A notable observation was the limitation in the coverage of different attacks present in multiple defensive methods. The observations made while examining different impersonation attacks and defensive methods showed that fusion models are effective protection method from impersonation attacks. Fusion models combine multiple methods for identifying and preventing impersonation attacks and therefore provide a large coverage of different attacks but their development and implementation demand time. A second notable protection method to stand out is human inspector which can be implemented quickly and is an effective defense method as long as the amount of data isn't large.

Keywords: Face recognition, impersonation attack, machine learning, defensive techniques

Sisällys

Kuvat

Taulukot

1 Johdanto

Kasvojentunnistusohjelmat ovat yleistyneet biometrisen tunnistautumisen keinona niin yksityisissä kuin julkisissa palveluissa. Kasvojentunnistusohjelmia käytetään esimerkiksi älypuhelinten lukituksen avaamisessa, rajatarkastuspisteillä tunnistautumisessa sekä yksityisalueiden ja -tilojen kulunvalvonassa. Maailman valtioista 60 prosenttia käyttää kasvojentunnistusta lentokentillään, 40 prosenttia kertoo osan työpaikkoistaan ottaneen kasvojentunnistuksen käyttöön ja 80 prosenttia valtioista kertoo maansa pankkien tai muiden finanssi-instituutioiden käyttävän kasvojentunnistusta [1].

Koska kasvojentunnistus on yleistynyt niin merkittävänä tunnistautumisen muotona, uhka rikollisten tai muiden vihamielisten tahojen sabotoivan kasvojentunnistusta päästäkseen käsiksi yksityisiin tietoihin tai valvottuihin alueisiin on kasvanut.

Kasvojentunnistuksen ohittamiseen on kaksi eri lähestymistapaa: toisen ihmisen imitointi (eng. impersonation) tai tunnistuksen välttö (eng. dodging). Tunnistuksen välttö koskee oman indentiteetin salaamista. Jos esimerkiksi julkiselle paikalle on asennettu kameroita tunnistamaan viranomaisten tarkkailulistoilla olevia henkilöitä, voi omaa yksityisyyttään suojata eri tunnistuksen välttämisen keinoin. Imitaatio saa kasvojentunnistusohjelman tunnistamaan ihmisen eri henkilöksi kuin hän on. Imitaatiohyökkäykset voidaan jakaa kahteen ryhmään: kohdistettuihin ja harkitsemattomiin hyökkäyksiin. Harkitsemattomassa hyökkäyksessä halutaan kasvojentunnistusohjelman tunnistavan hyökkääjä keneksi tahansa toiseksi henkilöksi.

Kohdistetussa hyökkäyksessä hyökkääjä haluaa kasvojentunnistusohjelman tunnistavan hänet tietyksi henkilöksi, kuten älypuhelimen omistajaksi.

Tämä tutkielma on kirjallisuuskatsaus, jossa pyritään löytämään erilaiset kasvojentunnistusta uhkaavat kohdistetut imitaatiohyökkäykset. Lisäksi arvioidaan näiden hyökkäysten tunnistamisen ja estämisen keinoja, sekä ehdotetaan uusia kehityskohteita, joilla kasvojentunnistus saadaan paremmin suojattua. Tutkielmassa vastataan seuraaviin tutkimuskysymyksiin:

1. Mihin kasvojentunnistukseen vaiheisiin voi toteuttaa kohdistettuja imitaatiohyökkäyksiä?

2. Mitä ratkaisuja hyökkäysten estoon on kehitetty?

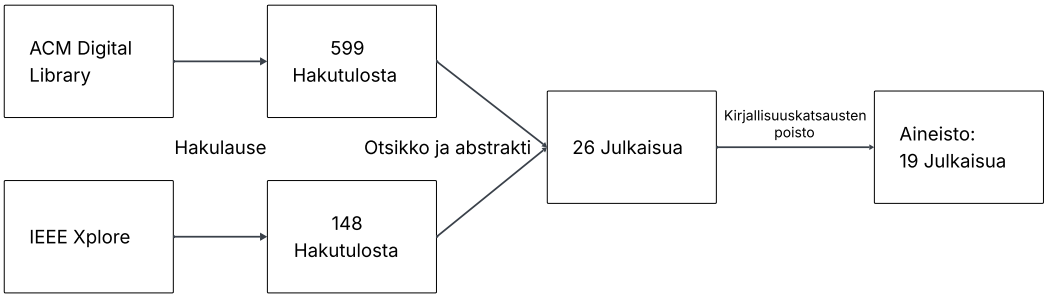
3. Mitä ratkaisuja on kannattava tutkia ja kehittää jatkossa?

Aiemmat tutkimukset kasvojentunnistusohjelmiin kohdistetuista hyökkäyksistä ovat käsitelleet yhteen kasvojentunnistusprosessin vaiheeseen kohdistuvia hyökkäyksiä kuten Poosideh ym. (2024) [2] tai ne ovat yleiskatsauksia kaikista hyökkäyksistä kuten Leyva, Gregory & Maple (2025) [3]. Näihin tutkimuksiin on kuulunut osana imitaatiohyökkäykset, mutta pelkästään imitaatiohyökkäyksiin keskittyvää aiempaa tutkimusta ei tiedonhaun ajankohtana löytynyt. Tämä tutkielma täydentää tätä vähemmän tarkasteltua kasvojentunnistukseen kohdistuvaa uhkaa.

Tutkielmaa varten suoritettiin tiedonhaku ACM Digital Library ja IEEE Xplore tietokannoista seuraavalla hakulauseella: ("facial recognition"OR "face recognition") AND impersona* AND (spoof* OR attack*).

Hakulauseessa on yhdistetty facial- ja face recognition, joilla saadaan kasvojentunnistusta käsitteleviä tuloksia. Impersona* hakutermillä saadaan imitaation liittyviä tuloksia ja spoof* OR attack* kohdistaa tulokset paremmin hyökkäyksiä käsitteleviin hakutuloksiin. Hakua ei ole muuten rajattu. ACM Digital Library tuotti hakulauseella 599 julkaisua ja IEEE Xplore 148 julkaisua.

Hakutuloksista tarkempaan tarkasteluun valittiin julkaisuja niiden otsikon ja ab-



Kuva 1.1: Aineistohaku

straktin mukaan. Otsikon ja abstraktin pohjalta tarkasteltiin oliko julkaisu aiheen kannalta merkittävä ja keskityttiinkö julkaisussa kokonaan tai merkittävästi kohdistettuihin imitaatiohyökkäyksiin. Vältettiin valitsemasta useampaa samaa aihetta käsittelevää julkaisua. Jos julkaisujen aihe oli sama, valittiin säilytettävä artikkeli sen merkittävyyden mukaan tutkielman aiheen ja tutkimuskysymysten kannalta. Tarkasteluun jäi 26 julkaisua, joista poistettiin kirjallisuuskatsaukset. Lopulliseen tarkasteluun jäi 19 julkaisua. Aineistojen haku ja valinta on esitetty kuvassa 1.1.

Luvussa 2 kerrotaan kasvojentunnistusteknologian toimintaperiaatteista ja käydään läpi kasvojentunnistusprosessin eri vaiheet. Luvussa 3 käydään läpi aineistohaulla löytyneitä erilaisia kohdistetun imitaatiohyökkäyksen toteutustapoja. Eri imitaatiohyökkäykset on jaettu neljään osa-alueeseen, joihin aineistojen esittämät hyökkäykset on luokiteltu. Aineistojen esittämien hyökkäysten jakauma eri osa-alueisiin esitetään taulukkossa 3.1. Hyökkäyksiä myös arvioidaan niiden luoman uhan suhteen. Arviointi perustuu hyökkäyksen toteutettavuuteen tosielämän tilanteissa ja kuinka paljon asiantuntijuutta kasvojentunnistusteknologiasta hyökkäyksen toteut-

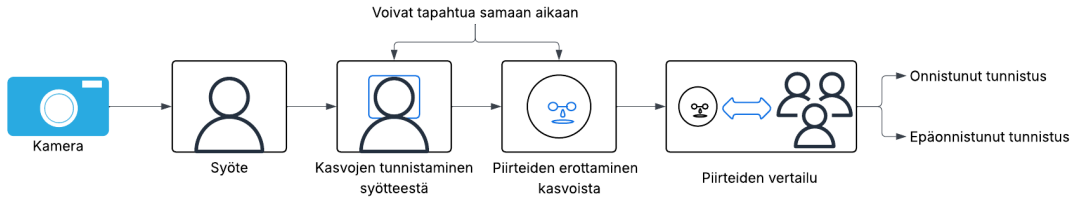
taminen vaatii. Luvussa 4 käydään läpi miten imitaatiohyökkäyksiä voidaan tunnistaa ja miten niiltä voi suojautua. Aineistossa ja muissa tutkimuksissa esitettyjä suojaus- ja tunnistuskeinoja arvioidaan niiden rahallisten kustannusten, käyttäänoton sujuvuuden ja sisältämien heikkouksien perusteella. Luvussa 5 ehdotetaan arvioinnin perusteella esille tulleita merkittäviä suojaustoimia, joilla saadaan kasvojentunnistuksen saadaan tehokkaasti suojattua. Luku 6 on yhteenveto tutkielmasta ja sen tuloksista.

2 Perehdytys kasvojentunnistukseen

Kasvojentunnistus on yksi tutkituimpia konenäön tehtäviä. Konenäön tavoitteena on saada kone näkemään ympäristönsä ihmisen tavoin. Konenäköön kuuluu useita tehtäviä, kuten kuvien tunnistus, luokittelu ja analysointi. Konenäkö on kehittynyt merkittävästi syväoppivien neuroverkkojen myötä. Neuroverkoista erityisesti konvoluutioneuroverkot (eng. convolution neural network) ovat yleisesti pohjana konenäön algoritmeihin [4].

Konvoluutioneuroverkot käsittelevät kuvaa sen pikseleistä koostettuna matriisina. Kuvasta tehty matriisi käydään läpi pienellä matriisilla, joka toimii suodattimena ja luo piirrekarttoja pienistä osista kuvaa. Kun piirrekarttoja käsitellään, malli erottaa kuvasta piirteitä automaattisesti. Malli erottaa ensin yksinkertaisia piirteitä kuten kaaria ja siirtyy niiden jälkeen isompiin kokonaisuuksiin kuten kasvoihin tai kukkiin. Konvoluutioneuroverkkon tarkempi toiminta ja rakenne riippuu sen arkkitehtuurista, joita on kehitetty useita [5].

Kasvojentunnistus kuuluu kuvantunnistustehtäviin. Kuvantunnistustehtävissä pyritään kuvasta erottamaan, rajaamaan ja tunnistamaan piirteitä. Kuvantunnistus koostuu kuvan esikäsittelystä, piirteiden erottamisesta ja luokittelusta. Esikäsittely parantaa tunnistamisen tarkkuutta poistamalla kuvasta häiriöitä. Piirteiden erottamisessa kuva muutetaan tietokoneelle käsiteltävään muotoon vektorikartaksi tai numeeriseksi esitykseksi. Luokittelussa kuva tunnistetaan saatujen piirteiden ja luokittelutehtävän perusteella [6].



Kuva 2.1: Kasvojentunnistusohjelman toiminta

Kasvojentunnistusta on sovellettu muun muassa identifioinnissa, kulunvalvonnassa, sekä ihmisen ja tietokoneen välisessä vuorovaikutuksessa. Automaattista kasvojentunnistusta on tutkittu jo 1970-luvulta alkaen, mutta 2000-luvulla koneoppimismenetelmien soveltaminen kasvojentunnistukseen toi sen osaksi arkea [7].

Muihin biometrisiin piirteisiin, kuten sormenjälkiin tai iiriskuviin, verrattuna kasvot ovat epätarkka tapa tunnistautua [7]. Kasvoilla biometrisenä tunnisteena on kuitenkin etunsa. Kasvojentunnistukseen käytettävät kuvat ovat helpompia ja nopeampia ottaa kuin sormenjäljet tai iiriskuvat. Lisäksi kasvojentunnistus vaatii vähemmän vuorovaikutusta käyttäjältä muihin edellä mainittuihin biometrisiin piirteisiin verrattuna. Kasvot ovat myös mielekkäämpiä verrata ja tarkastaa ihmisen toimesta esimerkiksi kulunvalvontatilanteessa verrattuna asiantuntijuutta vaativien sormenjälki- tai iiriskuvien vertailuun.

Ennen 2000-lukua ja koneoppimista, kasvojentunnistus perustui matemaattisiin malleihin kuten pääkomponenttianalyysiin (eng. principal component analysis). Kasvojentunnistus yleistyi arkiseen käyttöön kuitenkin koneoppimisen ja neuroverkkojen syväoppimisen tuoman tehokkuuden kautta.

2.1 Kasvojentunnistusohjelman toiminta

Automaattinen kasvojentunnistusprosessi voidaan jakaa neljään vaiheeseen: syöteen saaminen, kasvojen tunnistaminen syötteestä, piirteiden erottaminen tunnistetuista

kasvoista ja piirteiden vertailu. Toimintavaiheet on esitetty kuvassa 2.1.

Syöte on kasvojentunnistusohjelman saama kuva. Ohjelman toiminallisuudesta riippuen syöte voi olla osa videokuvaa, yksittäinen kuva tai 3D-kuva [4]. Syötteen muodosta riippumatta ensimmäinen vaihe ohjelmassa on saada syöte järjestelmään kytketyltä kameralta. Ohjelman kaikki vaiheet eivät välttämättä tapahdu samassa laitteessa, vaan esimerkiksi kamera voi lähettää syöteen verkon kautta toiselle koneelle seuraavia vaiheita varten.

Toisessa vaiheessa ohjelma tunnistaa onko syötteessä kasvot. Mikäli kasvoja ei tunnisteta, ohjelma päättyy. Jos kasvot tunnistetaan, ne voidaan rajata syötteestä ja siirtyä seuraavaan vaiheeseen.

Seuraavassa vaiheessa syötteestä rajatuista kasvoista erotetaan sen piirteet. Näitä piirteitä voivat olla silmien välinen etäisyys, nenän leveys ja leukaluun muoto [4]. Eri ohjelmat voivat erottaa eri piirteitä, mutta lopulta piirteet muutetaan numeeriseksi dataksi. Koska samoja piirteitä käytetään myös toisessa vaiheessa kasvojen tunnistamiseen kuvasta, voivat nämä vaiheet tapahtua samanaikaisesti.

Viimeisessä vaiheessa dataksi muutettuja piirteitä verrataan tallennettuihin piirteisiin ja etsitään vastaavuutta. Vertailua on kahdenlaista. Ensimmäinen on vahvistusvertailu (eng. verification), jossa syötettä verrataan yksiin talletettuihin kasvojen piirteisiin. Näin toimii esimerkiksi kännykän lukituksen avaaminen. Toinen on tunnistusvertailu (eng. identification), jossa etsitään mitä tahansa vastaavuutta kaikista järjestelmään talletetuista kasvoista. Tätä käytetään esimerkiksi kulunvalvonassa. Riippuen järjestelmän toteutuksesta voidaan molempia vertailuja käyttää samanaikaisesti [4].

Jotkin ohjelmat antavat vertailun päätteeksi samankaltaisuuspisteytyksen, joka kertoo kuinka lähellä syöteen piirteet olivat talletettuja piirteitä. Kun samankaltaisuuspisteet ovat tietyn rajan ylittäviä, ohjelma toteaa syöteen kasvojen vastaavan talletettuja kasvoja ja kasvot on tällöin tunnistettu. Vahvistusvertailua tekevät

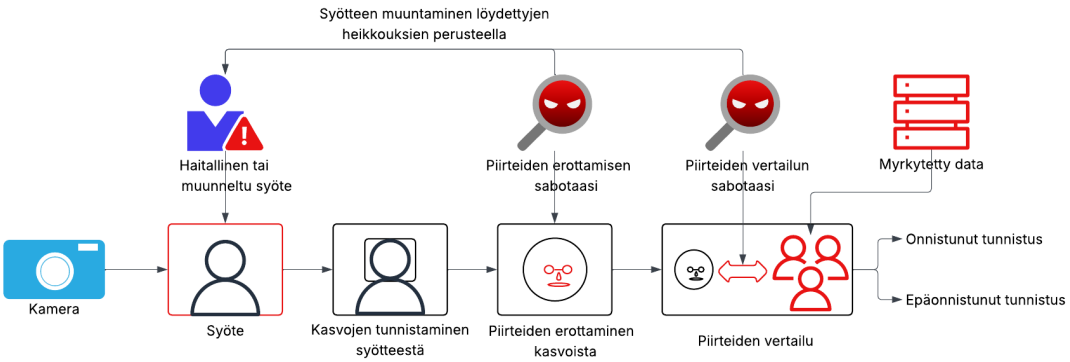
järjestelmät antavat yleensä samankaltaisuuspisteytyksen. Tunnistusvertailua käyttävät järjestelmät eivät palauta tietoa samankaltaisuudesta, vaan ne palauttavat tiedon onko kasvot tunnistettu vai ei [4].

3 Erilaiset imitaatiohyökkäykset

Imitaatiohyökkäykset voidaan jakaa neljään eri osa-alueeseen sen perusteella, mihin kasvojentunnistusprosessin vaiheeseen hyökkäys kohdistuu. Nämä osa-alueet ovat syöte, piirteidentunnistus, piirteiden vertailu ja ohjelman data. Osa-alueista piirteidentunnistus ja -vertailu ovat samat vaiheet, jotka esitettiin luvussa 2.1.

Kasvojen tunnistamista kuvasta ei tarkastella, koska tähän vaiheeseen kohdistuvat hyökkäykset keskittyvät estämään kasvojen tunnistamisen kuvasta. Imitaatiohyökkäyksissä halutaan kasvojen löytyvän syötteestä, joten tähän vaiheeseen kohdistuvat hyökkäykset eivät ole aiheen kannalta merkittäviä. Sen sijaan tarkastellaan syötteeseen kohdistuvia hyökkäyksiä. Nämä hyökkäykset pyrkivät onnistuneeseen imitaatioon pelkästään syötettä muokaamalla.

Ohjelman data kattaa kasvojen piirteiden vertailuun käytettävän talletetun da-



Kuva 3.1: Kasvojentunnistusohjelmaan kohdistuvat hyökkäykset

Taulukko 3.1: Kasvojentunnistuksen eri vaiheisiin kohdistuvien imitaatiohyökkäysten jakauma

	Hyökkäyksen toteutustapa			
Lähde	Haitallinen syöte	Piirteiden tunnistuksen sabotaasi	Piirteiden vertailun sabotaasi	Datan myrkytys
Cárabe ja Cermeño [8]		x		
Chen ym. [9]	x			
Cole ym. [10]				x
Farrokh Baroughi ym. [11]			x	
Gallardo-Cava ym. [12]	x			
He ym. [13]	x			
Kasichainula ym. [14]	x			
Li ym. [15]			x	
Lovisotto ym. [16]				x
Ma ym. [17]		x		
Petrelli ym. [18]	x			
Qin ym. [19]	x			
Sharif ym. [20]	x			
Shmelkin ym. [21]		x		
Tan ym. [22]		x		
Tariq ym. [23]	x			
Voth ym. [24]				x
Wang. ym. [25]	x			
Yan ym. [26]	x			

tan eli järjestelmään talletetut kasvojen piirteet, sekä kasvojentunnistusohjelman kouluttamiseen käytettävän datan. Koska yleisimmät ja tarkimmat kasvojentunnistusohjelmat perustuvat koneoppimisen menetelmiin, voidaan niiden koulutusdata myrkyttää ja näin luoda haavoittuvuus järjestelmään. Ohjelman datalla viitataan tässä luvussa sekä kasvojentunnistusjärjestelmään talletettuun dataan, että kasvojentunnistusohjelman käyttämään koulutusdataan.

Johdannossa 1 esitettiin tutkielman aineistojen hakuprosessi, jonka lopulliseen tarkastelluun jääneet artikkelit koottiin taulukoksi 3.1. Taulukko koottiin sen mukaan mihin imitaatiohyökkäyksen osa-alueeseen artikkelin esittämä hyökkäys tai hyökkäykset kuuluvat. Taulukosta nähdään, että imitaatiohyökkäyksiä voidaan koh-

distaa kaikkiin kasvojentunnistuksen vaiheisiin, jotka ovat imitaation kannalta merkittäviä.

Ennen erilaisten hyökkäysten läpikäyntiä avataan niihin liittyviä termejä. Musta laatikko viitta tilanteeseen, jossa kasvojentunnistusohjelmasta ei ole mitään tietoa hyökkääjällä. Harmaa laatikko viittaa tilanteeseen, jossa hyökkääjä saa kasvojentunnistusohjelmalta palautetta, kuten samankaltaisuuspisteytyksen tulokset. Valkoinen laatikko viittaa tilanteeseen, jossa hyökkääjä saa kasvojentunnistusohjelmalta palautetta ja hyökkääjällä on tiedossa ohjelman koulutusdata. Syväoppimista käyttävät kasvojentunnistusohjelmat sisältävät neuroverkkojen rakenteeseen kuuluvan piilotetun kerroksen. Koska piilotetun kerroksen toiminnasta ei ole tarkkaa tietoa kellään toimijalla, on valkoinen laatikko paras mahdollinen tieto, mikä hyökkääjällä voi olla.

3.1 Syötteeseen kohdistetut hyökkäykset

Syötteeseen kohdistetut imitaatiohyökkäykset ovat kaikkein yleisimpiä kuten taulukosta 3.1 nähdään. Tämä johtuu niiden helposta toteutettavuudesta ja toimivuudesta. Syötteeseen kohdistuville hyökkäyksille onkin annettu nimitys esityshyökkäys (eng. presentation attack).

Esityshyökkäys tarkoittaa syötteessä esiintyvän henkilön ulkonäköön vaikuttamista, joko fyysistä ulkomuotoa muuttamalla tai koneellisesti syötettä muokkaamalla. Tavoitteena on muuttaa oma ulkonäkö esittämään mahdollisimman hyvin toista henkilöä. Esityshyökkäys voidaan toteuttaa monella tapaa, kuten meikkaamalla [9][12], proteesi- tai lateksimaskeilla [12], kasvomaskeilla [13], silmälaseilla [20], kätketyillä häirintälaitteilla [25], sekä kasvoja peittävillä vaatekappaleilla [18]. Myös 3D syötettä käyttäviin kasvojentunnistus ohjelmiin voidaan kohdistaa fyysisiä esityshyökkäyksiä [26]. Erilaisilla maskeilla ja maskeerauksella toteutetuilla hyökkäyksillä on näistä hyökkäyksistä merkittävin onnistumisprosentti suhteessa toteuttamisen vaivattomuuteen [12]. Häirintälaitteet ja muut erikoistuneet hyökkäykset yltävät

puolestaan korkeampiin onnistumisprosentteihin mutta vaativat ymmärrystä, tietoa ja asiantuntijuutta kasvojentunnistusohjeman toiminnasta.

Koneellisessa esityshyökkäyksessä voidaan hyökkäyksen kohteena olevan henkilön kuvista lisätä irrotettuja paloja, kuten silmät tai nenä, hyökkääjän syötteeseen saaden ohjelma tunnistamaan hyökkääjän kohteena [19] tai muuttamalla suoraan kasvojen ulkonäköä [14]. Myös syväväärennöksillä (eng. deepfake) voidaan toteuttaa koneellinen esityshyökkäys [23]. Koneellisista esityshyökkäyksistä erityisesti syväväärennöksillä toteutetut hyökkäykset ovat yleisiä niiden helpon toteutettavuuden ja merkittävän onnistumisprosentin vuoksi. Kaupallisten kasvojentunnistusohjelmien syvähuijauksien tunnistustarkkuuden on osoitettu laskevan jopa 50 prosenttia, kun niitä testataan koulutusmateriaaliin kuulumattomilla syvähuijauksella tehdyillä syötteillä [27]. Syötteen muuttaminen kuvanmuokkausmetodeilla vaatii puolestaan tietoa kasvojentunnistusohjelman tunnistamisessa painotetuista piirteistä parhaan onnistumisprosentin saavuttamiseksi.

3.2 Piirteiden tunnistukseen kohdistetut hyökkäykset

Piirteiden tunnistukseen kohdistetut hyökkäykset sisältävät useassa tapauksessa myös syötteeseen vaikuttamista. Niitä kuitenkin yhdistää ensisijainen tavoite sabotoida piirteiden tunnistusta hyödyntämällä sen heikkouksia.

Sabotaasitarrat (eng. adversarial patch) ovat syötteeseen lisättäviä tarroja, joilla pyritään sabotoimaan syväoppimiseen perustuvien kasvojentunnistus ohjelmien piirteiden tunnistusta [17]. Tarrat voidaan luoda kasvojentunnistusohjelman koulutusdatan pohjalta tai ohjelman antaman palautteen perusteella. Kun tarra lisätään syötteeseen, ohjelma löytää siitä piirteitä, jotka vastaavat imitaation kohteena olevaa henkilöä.

Jos hyökkääjällä ei ole tietoa kohteen ulkonäöstä, voidaan kohteen kasvot luoda pelkästään piirteiden tunnistuksessa tarkasteltujen piirteiden ja samankaltaisuuspisteytyksen pohjalta. Näin luodut kasvot menevät läpi kasvojentunnistuksesta [21]. Samanlainen hyökkäys hyödyntää kasvojensulauttamisprosessin (eng. face embedding) tunnistamia piirteitä ja samankaltaisuuspisteytyksen tuloksia kasvojen uudelleen luomiseen [22].

Sabotaasitarrat ovat näistä hyökkäyksistä tehokkaimmat. Ne voidaan toteuttaa musta laatikko -tilanteissa [17] ja niillä on merkittävä onnistumisprosentti, mutta niiden yleistettävyys eri kasvojentunnistusohjelmien välillä on vielä heikkoa. Uudelleen luoduilla kasvoilla toteutetut hyökkäykset toimivat tehokkaasti vain, kun hyökkääjällä on tietoa kasvojentunnistusohjelmasta. Ne eivät ole helppoja toteuttaa tosielämän musta laatikko -tilanteissa ja vaativat hyökkääjältä asiantuntijuutta.

3.3 Piirteiden vertailuun kohdistetut hyökkäykset

Piirteiden vertailuun kohdistetut hyökkäykset tapahtuvat yleensä selvittämällä, minkä piirteiden tarkasteluun ohjelma keskittyy ja sabotoimalla ohjelman toimintaa vaikuttamalla näihin piirteisiin.

Muodonmuutoshyökkäys (eng. morphing attack) toimii muuntamalla kasvojentunnistusohjelman saamaa syötettä kunnes se läpäisee piirteiden vertailuvaiheen [28]. Muokkaus tapahtuu yleensä piirteiden vertailulta saadun palautteen perusteella. Muodonmuutoshyökkäys toimii myös tunnistusvertailua vastaan [8].

Hyökkäykset voivat kohdistua myös ohjelmakoodin heikkouksiin. Kuten luvussa 2.1 kerrottiin, eri ohjelmat ottavat syötteestä eri piirteitä. Histogrammeja käyttävissä kasvojentunnistus ohjelmissa esiintyy usein negatiivisen numeron bugi, jota hyväksikäyttämällä voidaan toteuttaa hyökkäys piirteiden vertailuun [11]. Hyökkäys toteutetaan esittämällä syötteessä sisältöä, joka vastaa imitaation kohteena olevan henkilön dataksi muutettuja piirteitä. Tällöin ohjelmassa aiheutuu negatiivisen nu-

meron bugi ja ohjelma tunnistaa hyökkääjän imitaation kohteena olevaksi henkilöksi.

Musta laatikko -tilanteissa hyökkäysten kohdistaminen piirteiden vertailua vastaan on hankalaa. Tällaisessa tilanteessa hyökkääjä voi käyttää valkoisen laatikon ohjelmaa, joka toteuttaa samaa toimminnallisuutta kuin mustan laatikon ohjelma. Valkoisen laatikon ohjelmalta saatuja tuloksia voidaan käyttää hyökkäyksessä mustan laatikon ohjelmaa vastaan. Sisarushyökkäyksessä (eng. sibling attack) syötettä kehitetään piirretunnistus (eng. attribute recognition) ohjelman tunnistamia piirteitä muokaamalla [15]. Tämän jälkeen syöte annetaan kasvojentunnistusohjelmalle. Syötettä muokataan kunnes se menee tunnistuksesta läpi.

Näistä hyökkäyksistä muodonmuutoshyökkäykset ovat tutkituimpia niiden helpon toteutettavuuden ja toimivuuden vuoksi. Muodonmuutoshyökkäykset on kuitenkin helppo tunnistaa heti, kun syötettä muutetaan useamman prosenttiyksikön verran [8], joten vain suorituskyvyltään tehokkaimmat muodonmuutoshyökkäykset ovat uhkia. Ohjelmakoodissa olevat heikkoudet luovat puolestaan ongelmia vain jos ne jäävät huomaamatta ja ovat yleensä korjattavissa, kun ne huomataan. Sisarushyökkäys luo puolestaan uniikin uhan, koska ainoa merkki hyökkäyksestä on kasvojentunnistusohjelman saama syöte. Sisarushyökkäyksen toteuttaminen vaatii kuitenkin asiantuntijuutta ja on täten epätodennäköinen tosielämän tilanteissa.

3.4 Ohjelman datan myrkytyshyökkäykset

Ohjelman datan myrkytyshyökkäyksillä pyritään lisäämään hyökkääjän myrkyttämää dataa ohjelmaan talletetun datan joukkoon tai laskemaan rajaa, jonka perusteella kasvot todetaan tunnistetuiksi. Kun myrkytetään talletettua dataa, hyökkääjä pyrkii lisäämään omat kasvonsa valtuutettujen käyttäjien joukkoon ja näin päästä läpi tunnistusprosessista. Voidaan myös myrkyttää yksittäisen talletetun kasvon piirteet, jolloin hyökkääjän on helpompi esiintyä kyseisenä henkilönä.

Myrkytetty data voidaan syöttää järjestelmään osana kohteen omia kuvia, kun rekisteröidään uutta käyttäjää [10] tai kun rekisteröity käyttäjä päivittää talletettuja kuviaan [16].

Ohjelman koulutusdataa myrkyttäessä pyritään joko laskemaan yleisellä tasolla ohjelman tunnistautumiseen vaadittavaa piirteiden samankaltaisuutta tai kohdistetuissa hyökkäyksissä luomaan hankalammin havaittava heikkous järjestelmään. Kohdistetussa koulutusdatan myrkytyshyökkäyksessä pyritään myrkyttämään tietty piirre tai tietyn henkilön kasvot ja näin luoda heikkous ohjelmaan [24].

Koulutusdatan myrkyttäminen on näistä hyökkäyksistä helpommin havaittavissa, koska koulutusdataa tarkastellaan paljon kasvojentunnistusohjelmaa kehittäessä. Kohteen syöttämän datan myrkyttäminen on puolestaan mieluisampi hyökkäystapa, koska se on vaikeampi havaita. Toisaalta se vaatii hyökkääjältä teknistä osaamista, sillä sen toteuttamiseksi on tehtävä väliintulohyökkäys, jotta käyttäjää saadaan huijattua ja silloinkin on luotettava käyttäjän huolimattomuuteen [10].

4 Imitaatiohyökkäyksiltä suojautuminen

Moneen eri imitaatiohyökkäykseen on jo kehitetty tunnistus- ja suojauskeinoja. Monet näistä ratkaisuista sisältävät erillisen toiminnallisuuden lisäämisen kasvojentunnistusprosessiin, jolloin tavoitteena on tunnistaa tietyn tyyppinen hyökkäys. Osa ratkaisuista on myös käytäntöjä, joita voidaan soveltaa useaa eri hyökkäystä vastaan.

Eri suojauskeinoja tarkastellessa ja arvioidessa on huomioitava missä kontekstissa hyökkäys tapahtuu. Jos hyökkäys vaatii asiantuntijuutta tai syvää ymmärrystä kasvojentunnistuksen toiminnasta, on huomattavasti epätodennäköisempää, että hyökkäys toteutettaisiin ja siltä suojautumista ei välttämättä tarvita. Vastaavasti, jos hyökkäystä ei voi toteuttaa julkisessa tilassa ohikulkijoiden tai esimerkiksi vartijan läsnäollessa, on kyseinen hyökkäys huomattavasti epätodennäköisempi uhka.

4.1 Syötteeseen kohdistuvien hyökkäysten tunnistaminen ja estäminen

Esityshyökkäyksien tunnistamiseen on kehitetty runsaasti erilaisia esityshyökkäyksen tunnistamisen metodeja [29][2]. Esityshyökkäyksen havaitseminen (eng. Presentation Attack Detection) kattaa monta eri tapaa tunnistaa esityshyökkäykset. Eri

tavat voidaan jakaa neljään alaluokkaan.

Tekstuurianalyysi (eng. Texture analysis), joka pyrkii erottamaan aidot kasvot valokuvista, näytöistä, maskeista ja muista kasvoja esittävistä asioista. Liikeanalyysi (eng. Motion analysis) tunnistaa pään liikeen. Se toimii erityisesti kaksiulotteisten (2D) hyökkäysten, eli valokuvien ja näyttöjen, tunnistamiseen. Avaruudellisten piirteiden -analyysi (eng. Spatial features analysis) tunnistaa kolmiulotteisia (3D) piirteitä syötteestä ja toimii video- ja syvähuijaushyökkäysten tunnistamiseen. Elonmerkkianalyysi (eng. Life sign analysis) tunnistaa syöteestä silmänräpäytykset, suun liikkeet, veren kierron ja kasvojen ilmeet. Sitä käytetään sekä 2D että 3D hyökkäysten tunnistukseen [2].

Tekstuurianalyysi tarjoaa kattavimman suojan erilaisilta esityshyökkäyksiltä, mutta vain korkean resoluution syötteillä. Elonmerkkianalyysi tarjoaa suppeamman kattavuuden eri hyökkäysten tunnistamisessa ja vaatii syötteen korkeaa resoluutiota, mutta on laskentatehokkuudessa tekstuurianalyysia nopeampi. Avaruudellisten piirteiden -analyysi ja liikeanalyysi eivät tarvitse korkean resoluution syötettä, mutta vaativat syötteen olevan videokuvaa parhaimman tunnistuskyvynsä saavuttamiseksi [29][2]. Kaikilla luokilla on vahvuutensa ja heikkoutensa, jotka perustuvat niiden laskennalliseen tehokkuuteen, vankkarakenteisuuteen, vaadittuun syötteen muotoon ja kykyyn yleistää tunnistusta eri hyökkäyksien välillä.

Alaluokkien ratkaisut sisältävät vaatimuksia kasvojentunnistusjärjestelmältä. Tekstuurianalyysi ja avaruudellisten piirteiden -analyysi tarvitsevat erikoiskameroita luotettavan toimivuuden takaamiseksi. Liikeanalyysi vaatii puolestaan käyttäjän yhteistyötä pään liikuttamisessa ohjeistetusti luotettavan toimivuuden takaamiseksi. Lisäksi alaluokkien eri metodit hyvin erikoistuneiden hyökkäysten tunnistamiseksi vaativat tarkaan rajattua koulutusmateriaalia, jota ei aina ole riittävästi olemassa tai ollenkaan saatavilla [2].

Esityshyökkäyksten toteuttamisesta ja tunnistamisesta on tehty monia havainto-

ja eri tutkimusten yhteydessä. Proteesi- ja lateksimaskeilla toteutettaviin hyökkäyksiin on todettu tarvittavan kohteen yhteistyötä maskin teossa, jos halutaan tehokas hyökkäys [12]. Erilaiset kasvoja peittävät esineet tai vaatekappaleet on tunnistettavissa ihmistarkastajan läsnäollessa [25]. Hyökkäyksien havaitsemiseen käytettävän mallin kouluttaminen sopivalla koulutusdatalla parantaa hyökkäyksen tunnistamisen todennäköisyyttä merkittävästi [25][23][2].

Syötteeseen kohdistuvien hyökkäysten suojaustoimien arviointi

Esityshyökkäyksien havaitsemisen alaluokkien metodit luovat yhdessä vahvan suojan hyökkäyksiltä, mutta ei ole käytännöllistä tai kannattavaa ottaa kaikkia näitä metodeja käyttöön. Siitä aiheutuisi suuria rahallisia kuluja, käytössä olevien järjestelmien uusimista sekä kasvojentunnistuksen helppokäyttöisyyden heikkenemistä. Vajaiden metodien käyttöönotosta saadaan kuitenkin vajaa suojaus [2].

Käytännöllisempää on ihmistarkastajan läsnäolo paikoissa, joissa kasvojentunnistus tapahtuu, kuten rajatarkastuspisteillä. Näissä tilanteissa esityshyökkäys on hankala toteuttaa koneellisesti, koska järjestelmään ei ole pääsyä. Fyysinen esityhyökkäys on myös hankala toteuttaa, sillä ihmisen on helppo tunnistaa erilaiset maskit ja kasvoja peittävät esineet. Ongelmakohdiksi jää kuitenkin meikki ja vähemmän huomiota herättävät asusteet, kuten silmälasit, joilla hyökkäys voidaan toteuttaa. Esimerkiksi meikkihyökkäykset voidaan tunnistaa lämpökameralla, mutta ongelma syntyy, jos aidolla käyttäjällä on runsas meikkaus ja järjestelmä tunnistaa hänet tästä syystä hyökkääjäksi. Meikkaushyökkäys voidaan siis tunnistaa koneellisin keinoin, mutta vain jos vaaditaan käyttäjältä meikkauksen poistoa. Muuten on hyväksyttävä heikkous järjestelmässä. Mitä vahvemmaksi hyökkäystentunnistus halutaan, sitä enemmän jouduttaan vaatimaan vastaavia toimia käyttäjältä. Tällöin kasvojentunnistuksen helppokäyttöisyys kärsii laskien sen miellekkyyttä tunnistautumisen keinona.

Useassa tutkimuksessa ([25][23][2][3]) on todettu tarve koulutusmateriaalille, jolla kasvojentunnistusohjelmat saadaan tunnistamaan hyökkäykset. Monet esityshyökkäyksien havaitsemisen alaluokkien metodit perustuvat syväoppimiseen. Niiltä kuitenkin puuttuu kunnollinen koulutus kaikkien erilaisten hyökkäyksien tunnistamiseen, koska koulutusmateriaalia ei ole näistä hyökkäyksistä tarjolla [2]. Koulutusmateriaalin kokoaminen kaikista eri esityshyökkäyksistä vaatii paljon aikaa ja resursseja, mutta valmistumisensa jälkeen materiaalia voidaan käyttää kaikkien kasvojentunnistusohjelmien hyökkäyksentunnistuksen tehostamiseen. Tällöin säilytetään myös kasvojentunnistuksen helppokäyttöisyys ja vähennetään tarvetta ihmistarkastajalle.

4.2 Piirteiden tunnistukseen kohdistuvien hyökkäysten tunnistaminen ja estäminen

Piirteiden tunnistukseen kohdistuvat hyökkäykset perustuvat joko kasvojentunnistusohjelmalta saatuun palautteeseen, kuten piirteidenvertailuhyökkäykset, tai tietoon ohjelman koulutusdatasta. Koska piirteiden tunnistukseen kohdistuvat hyökkäykset sisältävät usein syötteeseen vaikuttamista, on esityshyökkäyksen havaitsemisen keinot osittain sopivia näiden hyökkäysten tunnistamiseen.

Ensisijainen suojaustapa piirteiden tunnistukseen kohdistuvien hyökkäysten estoon on salata kaikki kasvojentunnistusohjelman tiedot. Koulutusdatan tulee olla salattu ja ohjelman toimintaperiaatteiden tulee olla salattu. Näin pystytään suojautua kaikilta hyökkäyksiltä, jotka perustuvat koulutusdatan tai toimintamallin heikkouksiin [3][17]. Toisin sanoen on kaikkia julkisesti saatavilla olevia koulutusmateriaaleja ja valmiita kasvojentunnistusohjelmia pidettävä epäluotettavina.

Muodonmuutoshyökkäyksien tunnistamiseen on kehitetty omia metodejaan, jotka perustuvat häiriön tunnistamiseen syötteessä. Näiden metodien ja elonmerkkianalyysin on todettu olevan tehokkaita tunnistamaan syötteen muuntamista sisältä-

vät hyökkäykset, mutta ne eivät ole täydellisiä [8][22]. Myös piirteiden tunnistukseen kohdistuvia hyökkäyksiä käsittelevissä tutkimuksissa on todettu koulutusmateriaalin olevan puutteellista esimerkiksi muodonmuutoshyökkäysten tunnistamisen parantamiseksi [8].

Piirteiden tunnistuksen suojauskeinojen arviointi

Erittäin toimiva tapa suojautua piirteidentunnistukseen kohdistuvilta hyökkäyksiltä on salata kaikki tieto kasvojentunnistusohjelmasta. Lisäksi kaikkia julkisesti saatavilla olevia kasvojentunnistusohjelmia, olivat ne valmiiksi koulutettuja tai eivät, tulisi pitää epäluotettavina aivan kuten julkista koulutusdataakin.

Mikäli näitä ohjeita seurataan täydellisesti voidaan nähdä kaksi mahdollista lopputulosta. Ensimmäinen mahdollisuus on kaikkien kasvojentunnistusta käyttävien organisaatioiden tarvitsevan oman mallinsa, jonka he kouluttavat omalla koulutusdatallaan. Tämä vaatisi investointeja kaikilta kasvojentunnistusta käyttäviltä tahoilta ja lopputuloksena nähtäisiin useita pienellä määrällä dataa koulutettuja malleja. Toinen mahdollisuus on, että kaikki kasvojentunnistusta käyttävät tahot siirtyvät käyttämään vain muutamaa mallia, joiden tiedot on täysin salattuja. Tällöin yksikin heikkous laajasti käytetyssä mallissa olisi heikkous useassa järjestelmässä. Lisäksi organisaatiot tulisivat riippuvaiseksi näistä muutamasta mallista, jolloin mallin olessa poissa toiminassa kaikki sitä käyttävät palvelut ovat myös poissa toiminnasta.

Suurin ongelma, joka koskee molempia esitettyjä tilanteita on kuitenkin kaiken nykyisen datan ja tiedon hylkääminen. Jotta suojauskeino olisi tehokas, tulisi mallin toimintaperiaatteiden ja koulutusdatan olla täysin salaiset, eli täysin uudet. Ei siis voida hyödyntää mitään olemassa olevaa. Tästä syystä kaiken tiedon salaaminen on tehokas, mutta tosielämässä huono suojauskeino.

Suojauskeinoja, jotka tunnistavat syötteen muuntamisen, koskee samat aliluvussa 4.1 esitetyt ongelmat.

4.3 Piirteiden vertailuun kohdistuvien hyökkäysten tunnistaminen ja estäminen

Kaikki piirteiden vertailuun kohdistuvat hyökkäkset vaativat asiantuntijuutta. Hyökkäyksissä kasvojentunnistusohjelman saamaa syötettä muunnetaan piirteiden vertailusta saadun samankaltaisuuspisteytyksen tai vastaavan palautteen perusteella. Lisäksi monet näistä hyökkäyksistä toteutetaan synteettisellä datalla [3] eli syöte on luotu hyökkääjän toimesta ohjelmalta saadun palautteen perusteella.

Kuten piirteiden tunnistamisessa, on piirteiden vertailun toiminan salaaminen hyvä suojauskeino. Mustan laatikon ohjelmat eivät kuitenkaan anna täydellistä suojaa kaikkia hyökkäyksiä vastaan [15].

Kasvojentunnistusohjelmien kouluttaminen synteettisten syötteiden tunnistamiseksi on myös toimiva ratkaisu [15][3]. Myös erilaiset häiriöidenvähennysmenetelmät (eng. de-noise) auttavat poistamaan syötteestä hyökkääjän tekemiä muutoksia [15]. Kouluttaminen on näistä suojaustavoista parempi, koska se on paremmin toteutettavissa useisiin eri kasvojentunnistusjärjestelmiin. Häiriöidenvähennysmenetelmät sopivat järjestelmiin, joissa syötettä muutetaan muutenkin runsaasti ennen piirteiden tunnistusta. Häiriöidenvähennysmenetelmien ei kuitenkaan haluta muuttavan syötettä liian paljon, koska tämä laskee tunnistustarkkuutta [25], joten niiden tehokkuus hyökkäysten estossa on rajallinen.

Piirteiden vertailuun kohdistuvien hyökkäysten suojauskeinojen arviointi

Piirteiden vertailun toimintaperiaatteiden pitäminen salassa ja kaiken mahdollisen ohjelman antaman palautteen salassapito on toimiva suojauskeino, mutta ei täydellinen ratkaisu [15].

Kouluttaminen ja häiriöidenvähennysmenetelmät ovat puolestaan lupaavampia suojauskeinoja, koska tällöin tunnistetaan ja saadaan poistettua syötteestä hyökkää-

jän tekemiä muutoksia. Jos häiriöidenvähennysmenetelmät ja mallin kouluttaminen haitallisten syötteiden tunnistamiseksi yhdistetään, saadaan tehokas suoja [15]. Ongelmana on kuitenkin sopivan koulutusmateriaalin puute ja häiriönvähennyksen rajallinen tehokkuus suojauskeinona.

On hyvä huomioida, että piirteiden vertailuun kohdistuvissa hyökkäyksissä syötteellä on merkittävä rooli. Näin ollen hyvä suojaus haitallisia syötteitä vastaan takaa hyvän suojan piirteiden vertailulle.

4.4 Ohjelman datan myrkytyshyökkäysten tunnistaminen ja estäminen

Myrkytyshyökkäyksien tunnistamisessa on suuri merkitys sillä onko suojaajalla pääsy kasvojentunnistusohjelman dataan. Arkikäytössä olevat kasvojentunnistusohjelmat ovat suurimaksi osaksi kolmansien osapuolten halussa, jolloin suojausta toteuttavalla taholla ei ole pääsyä ohjelmaan [30].

Yksi metodi myrkytetyn datan tunnistamiseen on syväoppimista hyödyntävä datan tarkastaja, joka on erillään kasvojentunnistusjärjestelmästä. Hyökkääjän on vaikea luoda sopivaa myrkytytettyä dataa, joka huijasi kasvojentunnistusjärjestelmää, sekä erillistä tarkastajaa [10].

Vaihtoehtoinen ratkaisu erillisen tarkastajan sijaan on kouluttaa kasvojentunnistusohjelma tunnistamaan myrkytetty data syötteen keskiön perusteella. Koulutus voi vaatia kohdistettua dataa jokaisesta erilaisesta myrkytyshyökkäyksestä [16], mutta koulutuksen jälkeen ohjelma on tarkempi tunnistamaan hyökkäyksiä kuin erillinen datan tarkastaja. Tarkastaja tarjoaa kuitenkin paremman suojan, jos hyökkääjällä on täysi ymmärrys kasvojentunnistusjärjestelmästä.

Mikäli on syytä epäillä kasvojentunnistusohjelman olevan jo myrkytetty, voidaan kouluttaa uusi malli samalla koulutusdatalla kuin alkuperäinen malli. Mikäli uuden

mallin oppimat painot eroavat merkittävästi alkuperäisen mallin oppimien painojen arvoista, voidaan malli todeta myrkytetyksi [30]. Mikäli malli on tunnistettu myrkytetyksi ja kyseinen myrkytys laskee tunnistautumiseen vaadittua rajaa, voidaan malli kouluttaa unohtamaan heikkous [30].

Myrkytyshyökkäysten suojauskeinojen arviointi

Suurin haaste myrkytyshyökkäyksien estämisessä on estetty tai vain osittainen pääsy kasvojentunnistusohjelman koulutusdataan ja talletettuun dataan. Näin voi käydä, kun ohjelman tarjoaa jokin kolmas osapuoli, jonka täytyy muunmuassa yksityisyyden suojan takia evätä pääsy ohjelman dataan. Tällöin palvelun tarjoajan on suostuttava lisäämään suojauskeino järjestelmään.

Vastaavasti ohjelman olessa kolmannen osapuolen halussa, ei ole tietoa onko ohjelman koulutukseen käytetty epäluotettavaa julkista koulutusmateriaalia. Mikäli koulutusmateriaali on myrkytetty, ei uuden mallin kouluttaminen samalla datalla paljasta myrkytystä.

Kasvojentunnistusohjelman kouluttaminen tunnistamaan hyökkäykset ilman lisättyjä toiminnallisuuksia vaatii jälleen koulutusdataa. Tätä dataa ei vältämättä ole tarjolla ja koulutus yhtä myrkytyshyökkäystä vastaan ei takaa suojaa muilta myrkytyshyökkäyksiltä [16].

Myrkytyshyökkäyksiltä suojautumiseen tarvitaan yhteistyötä kasvojentunnistusohjelman omistavan tahon kanssa. Tätä yhteistyötä voi estää erilaiset lakitekniset syyt, mutta sen onnistuessa voidaan myrkytyshyökkäykset tunnistaa tehokkaasti.

5 Pohdinta

Erilaisia imitaatiohyökkäyksiä läpikäydessä tuli ilmi, että myrkytyshyökkäyksiä lukuun ottamatta kaikki imitaatiohyökkäykset muuttavat syötettä osana hyökkäystä. Jos kasvojentunnistusohjelma tunnistaa kaikki haitalliset syötteet, saadaan suoja lähes kaikilta imitaatiohyökkäyksiltä. Lisäksi on huomioitava, että erilaisten ehdotettujen ja olemassa olevien imitaatiohyökkäysten tunnistus- ja suojauskeinojen on osoitettu olevan tehokkaita tietyn tyyppisiä hyökkäyksiä vastaan, mutta yksikään yksittäinen ratkaisu ei pysty takaamaan täyttä suojaa kaikilta hyökkäyksiltä.

Näin ollen tehokas suojaus imitaatiohyökkäyksiltä saadaan aiemmissa tutkimuksissa ehdotetuilla fuusiomalleilla, jotka yhdistävät useita esityshyökkäyksen tunnistamisen keinoja. Erityisesti tekstuuri- ja elonmerkkianalyysin yhdistäminen yhdeksi tunnistimeksi luo kattavan suojan eri hyökkäyksiltä [2]. Fuusiomalliin voidaan lisätä myös syväoppimiseen perustuva tarkastaja, sillä Cole, Newman & Lin (2022) [10] ovat osoittaneet kahden erillisen syväoppivan neuroverkon samanaikaisen huijaamisen yhdellä syötteellä olevan erittäin vaikeaa. Neuroverkot olisivat kasvojentunnistusohjelma ja fuusiomalli.

Fuusiomallit ovat tehokkaammpia kuin yksittäiset esityshyökkäyksen tunnistamisen metodit, mutta niiden kehitystä hidastaa sama ongelma eli koulutusmateriaalin puute. Kuten aiemmin todettiin, hyökkäystentunnistuskoulutukseen ei ole olemassa tai ei ole riittävästi koulutusmateriaalia [2]. Ongelman ratkaisemista hankaloittaa se, ettei yhden hyökkäyksen tunnistamiseen käytettävä koulutusmateriaali takaa suojaa

muilta hyökkäyksiltä. Vaikka koulutusmateriaalia saataisiin kaikkien tällä hetkellä tiedossa olevien hyökkäysten tunnistamiseen, ei todennäköisesti kestäisi kauan, ennen kuin uusi hyökkäys löydettäisiin. Lisäksi pelkkä koulutus hyökkäyksen tunnistamiseksi ei takaa täydellistä suojaa kyseiseltä hyökkäykseltä.

Koska koulutusmateriaalin luominen vaatii huomattavasti aikaa ja rahaa, on tämän tutkielman havaintojen perusteella järkevämpää keskittää hyökkäystentunnistuskoulutus vain kaikkein yleisimpiin ja helpoimmin toteutettaviin imitaatiohyökkäyksiin. Kannattaa keskittyä esimerkiksi syvähuijausta hyödyntäviin esityshyökkäyksiin, joiden määrä on ollut kasvussa [31]. Syvähuijaushyökkäysten lisäksi tulisi tunnistaa muut yleisimmät imitaatiohyökkäykset ja luoda koulutusmateriaalia näiden hyökkäysten tunnistamiseksi.

Koska fuusiomallien kehittäminen ja kouluttaminen vaatii aikaa, tarvitaan nopeasti käyttöönotettavia suojatoimia kehittämisen ajaksi. Ihmistarkastaja on yksi helposti käyttöönotettava ja tehokas suojauskeino. Usea aineiston tutkimus totesi ihmistarkastajan tekevän hyökkäyksestä erittäin hankalan toteuttaa. Lisäksi tässä tutkielmassa esille nousseiden havaintojen perusteella ihmistarkastaja on sopiva usean erilaisen hyökkäyksen tunnistamiseen. Myös niiden aineistossa esitettyjen hyökkäysten tunnistamisessa, joita käsittelevissä tutkimuksissa ei tuotu esille ihmistarkastajaa ollenkaan.

Esimerkiksi rajatarkastuspisteellä ihmistarkastaja voi pitää silmällä erikoisia pääsä tai kasvoilla olevia asusteita ja ohjata näitä käyttävät henkilöt ihmisen suorittamaan tarkastukseen. Sähköisissä järjestelmissä ihminen voi tarkistaa minkälaisia syötteitä kasvojentunnistusohjelma saa. Lähes kaikki syötettä muokkaavat hyökkäykset ovat helposti ihmiselle tunnistettavia [2].

Ihmistarkastaja soveltuu myös myrkytyshyökkäyksien tunnistamiseen, vaikka ne eivät muuta syötettä. Mikäli kasvojentunnistusjärjestelmä tallettaa käyttäjien kasvojen piirteet myöhempää tunnistusta varten, ihmistarkastajan kannattaa käydä

säännöllisin väliajoin läpi talletettu data. Pelkästä piirredatasta on hankala tunnistaa myrkytettyä dataa, mutta mikäli järjestelmä tallettaa kuvia, on ihmisen jälleen helppo tunnistaa myrkytettyt kuvat. Vastaavasti ihminen voi käydä läpi kasvojentunnistusohjelman koulutusmateriaalin, jos on syytä epäillä sitä myrkytetyksi.

Ihmistarkastaja ei takaa täydellistä suojaa imitaatiohyökkäyksiltä, mutta on tehokas tapa tunnistaa hyökkäyksiä, jotka ovat koneellisesti hankalampia havaita. Ihmistarkastajalla on kuitenkin merkittävä haaste, joka on datan määrä. Koulutusmateriaalit kasvojentunnistukselle sisältävät miljoonia kuvia [7]. Lisäksi kasvojentunnistusohjelmaan on voitu tallettaa piirredataa miljoonista ihmisistä. Kun tähän lisätään kasvojentunnistusohjelmien päivittäinen käyttäjämäärä, on ihmistarkastajan työmäärä todella suuri. Datan määrän kasvaessa myös riski ihmisen tekemästä virheestä sitä tarkistaessa kasvaa, jolloin ihmistarkastajan tehokkuus suojauskeinona laskee.

Myrkytyshyökkäyksien tunnistamiseen ja estämiseen tarvitaan kasvojentunnistusohjelmia tarjoavien kolmansien osapuolten yhteistyötä. Yhteistyötä ja esitettyjen suojauskeinojen käyttöönottoa estää erityisesti lakitekniset syyt, jotka estävät talletetun datan tarkastelun. Sopivalla lainsäädännöllä voidaan poistaa tämä este ja taata parempi suoja myrkytyshyökkäyksiltä.

6 Yhteenveto

Kasvojentunnistusta on alettu käyttää yhä enemmän biometrisen tunnistautumisen keinona. Kasvojentunnistuksen käyttö kulunvalvonnassa, tunnistautumisessa ja yksityislaitteiden lukitusten avaamisessa on tehnyt siihen kohdistuvista hyökkäyksistä merkittävän riskitekijän.

Tässä tutkielmassa tarkasteltiin kohdistettuja imitaatiohyökkäyksiä uhkana kasvojentunnistukselle. Tarkasteltiin mitä eri tapoja kohdistetun imitaatiohyökkäyksen toteuttamiseen on olemassa. Tutkielmaa varten kerätyn aineiston artikkeleissa esitetyt hyökkäykset jaettiin neljään osa-alueeseen. Jako esitettään taulukossa 3.1, josta nähdään kohdistettujen imitaatiohyökkäysten olevan mahdollisia toteuttaa kaikkiin imitaation kannalta merkittäviin kasvojentunnistusprosessin vaiheisiin. Ainestossa esitetyt imitaatiohyökkäykset käytiin läpi ja tuotiin esille niiden toteuttettavuus tosielämän tilanteessa ja niiden toteuttamisen vaatima asiantuntijuus kasvojentunnistusteknologiasta

Aineistossa ja aiemmissa tutkimuksissa esitettyjä imitaatiohyökkäyksen suojauskeinoja käytiin läpi ja arvioitiin niitä rahallisten kustannusten, käyttäänoton sujuvuuden ja niiden sisältämien heikkouksien perusteella. Kun eri suojaus- ja tunnistuskeinoja läpikäytiin, huomattiin monien ehdotettujen ratkaisujen olevan hankalia toteuttaa niiden resurssivaatimusten takia ja lähes kaikkien ratkaisujen tarjoavan suojaa vain yhden lajin hyökkäyksiltä. Esille nousi erityisesti hyökkäysten tunnistamiseen tarvittavan koulutusmateriaalin puute. Jos koulutusmateriaalia saadan luo-

tua pelkästään yleisimmistä imitaatiohyökkäyksistä, voidaan kasvojentunnistusjärjestelmälle kouluttaa vahvempi suoja suurimpia uhkia vastaan. Tämän toteuttaminen kuitenkin vaatisi aikaa ja resursseja.

Aikaisempien tutkimusten ja tutkielmassa tehtyjen havaintojen perusteella todettiin fuusiomallien olevan erittäin tehokkaita kohdistettujen imitaatiohyökkäysten tunnistus- ja suojauskeinoja, joita kannattaa jatkossa tutkia ja kehittää. Erityisesti haitallisia syötteitä tunnistavan fuusiomallin todettiin kattavan suuren joukon eri hyökkäyksiä. Fuusiomallien kehittämistä hidastavan koulutusmateriaalin puutteen täydentämisen ajaksi suositeltiin nopeasti käyttöön otettavana suojauskeinona ihmistarkastajaa. Aiemmissa tutkimuksissa esille tuodun ja tässä tutkielmassa korostettun ihmistarkastajan tuominen osaksi kasvojentunnistusjärjestelmää on tehokas, yksinkertainen ja nopeasti käyttöönotettava tapa hyökkäysten tunnistamiseen. Ihmistarkastajan heikkous on kuitenkin kasvojentunnistusjärjestelmien käsittelemän datan valtava määrä, joka kasvattaa riskiä ihmisen tekemästä virheestä.

Lähdeluettelo

[1] Calvello, "32 Facial Recognition Statistics to Know in 2023", 2023. url: https://www.g2.com/articles/facial-recognition-statistics.

[2] M. Pooshideh et al., "Presentation Attack Detection: A Systematic Literature Review", *ACM Comput. Surv.*, vol. 57, nro 1, lokakuu 2024, ISSN: 0360-0300. DOI: 10.1145/3687264. url: https://doi.org/10.1145/3687264.

[3] R. Leyva, E. Gregory ja C. Maple, "Attack Vectors for Face Recognition Systems: A Comprehensive Review", *ACM Comput. Surv.*, vol. 58, nro 1, syyskuu 2025, ISSN: 0360-0300. DOI: 10.1145/3736753. url: https://doi.org/10.1145/3736753.

[4] L. Qinjun, C. Tianwei, Z. Yan, W. Yuying et al., "Facial recognition technology: a comprehensive overview", *Academic journal of computing & information science*, vol. 6, nro 7, s. 15–26, 2023.

[5] D. Bhatt et al., "CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope", *Electronics*, vol. 10, nro 20, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10202470. url: https://www.mdpi.com/2079-9292/10/20/2470.

[6] Y. Li, "Research and Application of Deep Learning in Image Recognition", teoksessa *2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications (ICPECA)*, 2022, s. 994–999. DOI: 10.1109/ICPECA53709.2022.9718847.

[7] I. Adjabi, A. Ouahabi, A. Benzaoui ja A. Taleb-Ahmed, "Past, Present, and Future of Face Recognition: A Review", *Electronics*, vol. 9, nro 8, 2020, ISSN: 2079-9292. DOI: 10.3390/electronics9081188. url: https://www.mdpi.com/2079-9292/9/8/1188.

[8] L. Cárabe ja E. Cermeño, "Stegano-Morphing: Concealing Attacks on Face Identification Algorithms", *IEEE Access*, vol. 9, s. 100 851–100 867, 2021. DOI: 10.1109/ACCESS.2021.3088786.

[9] C. Chen, A. Dantcheva, T. Swearingen ja A. Ross, "Spoofing faces using makeup: An investigative study", teoksessa *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2017, s. 1–8. DOI: 10.1109/ISBA.2017.7947686.

[10] D. Cole, S. Newman ja D. Lin, "A New Facial Authentication Pitfall and Remedy in Web Services", *IEEE Transactions on Dependable and Secure Computing*, vol. 19, nro 4, s. 2635–2647, 2022. DOI: 10.1109/TDSC.2021.3067794.

[11] A. Farrokh Baroughi, S. Craver ja M. F. Mohsin, "A Negative Number Vulnerability for Histogram-based Face Recognition Systems", teoksessa *Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security*, sarja IH&MMSec '15, Portland, Oregon, USA: Association for Computing Machinery, 2015, s. 155–160, ISBN: 9781450335874. DOI: 10.1145/2756601.2756617. url: https://doi.org/10.1145/2756601.2756617.

[12] R. Gallardo-Cava, D. Ortega-Delcampo, J. Guillen-Garcia, D. Palacios-Alonso ja C. Conde, "Creating Realistic Presentation Attacks for Facial Impersonation Step-by-Step", *IEEE Access*, vol. 11, s. 109 257–109 266, 2023. DOI: 10.1109/ACCESS.2023.3313094.

[13] C. He et al., "MysticMask: Adversarial Mask for Impersonation Attack Against Face Recognition Systems", teoksessa *2024 IEEE International Conference on*

Multimedia and Expo (ICME), 2024, s. 1–6. DOI: 10.1109/ICME57554.2024.10687792.

[14] K. Kasichainula, H. Mansourifar ja W. Shi, "RAF: Recursive Adversarial Attacks on Face Recognition Using Extremely Limited Queries", teoksessa *2022 IEEE International Conference on Big Data (Big Data)*, 2022, s. 1550–1555. DOI: 10.1109/BigData55660.2022.10020849.

[15] Z. Li et al., "Sibling-Attack: Rethinking Transferable Adversarial Attacks against Face Recognition", teoksessa *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, s. 24 626–24 637. DOI: 10.1109/CVPR52729.2023.02359.

[16] G. Lovisotto, S. Eberz ja I. Martinovic, "Biometric Backdoors: A Poisoning Attack Against Unsupervised Template Updating", teoksessa *2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2020, s. 184–197. DOI: 10.1109/EuroSP48549.2020.00020.

[17] H. Ma, K. Xu, X. Jiang, Z. Zhao ja T. Sun, "Transferable Black-Box Attack Against Face Recognition With Spatial Mutable Adversarial Patch", *IEEE Transactions on Information Forensics and Security*, vol. 18, s. 5636–5650, 2023. DOI: 10.1109/TIFS.2023.3310352.

[18] D. Petrelli, N. Dulake, G. Molinari ja F. Ciravegna, "Design-driven Deception of Face Recognition: An Empirical Study", *ACM Trans. Comput.-Hum. Interact.*, lokakuu 2025, Just Accepted, ISSN: 1073-0516. DOI: 10.1145/3769675. url: https://doi.org/10.1145/3769675.

[19] L. Qin, F. Peng, M. Long, R. Ramachandra ja C. Busch, "Vulnerabilities of Unattended Face Verification Systems to Facial Components-based Presentation Attacks: An Empirical Study", *ACM Trans. Priv. Secur.*, vol. 25,

nro 1, marraskuu 2021, ISSN: 2471-2566. DOI: 10.1145/3491199. url: https://doi.org/10.1145/3491199.

[20] M. Sharif, S. Bhagavatula, L. Bauer ja M. K. Reiter, "Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition", teoksessa *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, sarja CCS '16, Vienna, Austria: Association for Computing Machinery, 2016, s. 1528–1540, ISBN: 9781450341394. DOI: 10.1145/2976749.2978392. url: https://doi.org/10.1145/2976749.2978392.

[21] R. Shmelkin, T. Friedlander ja L. Wolf, "Generating Master Faces for Dictionary Attacks with a Network-Assisted Latent Space Evolution", teoksessa *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 2021, s. 01–08. DOI: 10.1109/FG52635.2021.9666968.

[22] M. Tan, Z. Zhou ja Z. Li, "The Many-faced God: Attacking Face Verification System with Embedding and Image Recovery", teoksessa *Proceedings of the 37th Annual Computer Security Applications Conference*, sarja ACSAC '21, Virtual Event, USA: Association for Computing Machinery, 2021, s. 17–30, ISBN: 9781450385794. DOI: 10.1145/3485832.3485840. url: https://doi.org/10.1145/3485832.3485840.

[23] S. Tariq, S. Jeon ja S. S. Woo, "Am I a Real or Fake Celebrity? Evaluating Face Recognition and Verification APIs under Deepfake Impersonation Attack", teoksessa *Proceedings of the ACM Web Conference 2022*, sarja WWW '22, Virtual Event, Lyon, France: Association for Computing Machinery, 2022, s. 512–523, ISBN: 9781450390965. DOI: 10.1145/3485447.3512212. url: https://doi.org/10.1145/3485447.3512212.

[24] D. Voth, L. Dane, J. Grebe, S. Peitz ja P. Terhörst, "Effective Backdoor Learning on Open-Set Face Recognition Systems", teoksessa *2025 IEEE/CVF Win-*

ter Conference on Applications of Computer Vision (WACV), 2025, s. 1027–1039. DOI: 10.1109/WACV61041.2025.00109.

[25] Y. Wang, Z. Liu, B. Luo, R. Hui ja F. Li, "The Invisible Polyjuice Potion: an Effective Physical Adversarial Attack against Face Recognition", teoksessa *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, sarja CCS '24, Salt Lake City, UT, USA: Association for Computing Machinery, 2024, s. 3346–3360, ISBN: 9798400706363. DOI: 10.1145/3658644.3670382. url: https://doi.org/10.1145/3658644.3670382.

[26] S. Yan, H. Wen, S. Chang, H. Zhu ja L. Zhou, "Fooling 3D Face Recognition with One Single 2D Image", teoksessa *Proceedings of the 32nd ACM International Conference on Multimedia*, sarja MM '24, Melbourne VIC, Australia: Association for Computing Machinery, 2024, s. 4043–4052, ISBN: 9798400706868. DOI: 10.1145/3664647.3680840. url: https://doi.org/10.1145/3664647.3680840.

[27] S. Kilany ja A. Mahfouz, "A comprehensive survey of deep face verification systems adversarial attacks and defense strategies", *Scientific Reports*, vol. 15, nro 1, s. 30 861, 2025.

[28] U. Scherhag et al., "Biometric Systems under Morphing Attacks: Assessment of Morphing Techniques and Vulnerability Reporting", teoksessa *2017 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 2017, s. 1–7. DOI: 10.23919/BIOSIG.2017.8053499.

[29] R. Ramachandra ja C. Busch, "Presentation Attack Detection Methods for Face Recognition Systems: A Comprehensive Survey", *ACM Comput. Surv.*, vol. 50, nro 1, maaliskuu 2017, ISSN: 0360-0300. DOI: 10.1145/3038924. url: https://doi.org/10.1145/3038924.

[30] Q. L. Roux, E. Bourbao, Y. Teglia ja K. Kallas, ”A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems”, *IEEE Access*, vol. 12, s. 47 433–47 468, 2024. DOI: 10.1109/ACCESS.2024.3382584.

[31] Khalil, ”Deepfake Statistics 2025: AI Fraud Data & Trends”, 2025. url: https://deepstrike.io/blog/deepfake-statistics-2025.