Софийски университет „Св. Кл. Охридски"

Факултет по математика и информатика

Катедра „Компютърна информатика"

ДИПЛОМНА РАБОТА

на тема

„ Система за засичане на характерни точки на лицето оптимизирана за работа в реално време на платформа за вградени системи"

Дипломант: Никола Ясенов Божинов

Магистърска програма: Изкуствен интелект

Факултетен номер: 26133

Научен ръководител:

д-р. Мартин Върбанов

Консултант:

проф. д-р Мария Нишева

София, 2021 г.

Съдържание

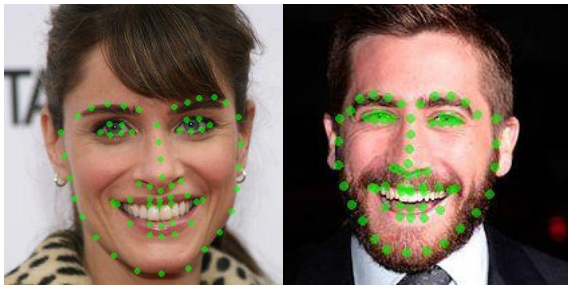
1. Увод

1.1 Актуалност на проблема и мотивация

Характерните точки на лицето (facial landmarks) описват основните части на лицето - очи, уста, нос, контури на лицето и се използват като основа за множество задачи, като например лицево разпознаване, четене по устните, разчитане на изражението на лицето или проследяване на погледа. Успехът на тези задачи е силно зависим от точността на локализиране на използваните от тях характерни точки на лицето. Тези задачи все по-широко се използват за цели като защита на мобилни устройства, контрол на физическия достъп до офиси и помещения, разпознаване на умората в шофьори и други, които изискват решаването им в реално време на крайните устройства. Съществуват редица хардуерни платформи, проектирани за изпълняване на задачи чрез изкуствен интелект, и в частност машинно самообучение, като част от вградени системи, но тяхната производителност все още значително изостава от тази на съвременните настолни компютри, изисквайки допълнителна оптимизация на изпълняваните софтуерни системи, за да работят в реално време на съответния специфичен хардуер.

Лицето играе важна роля във визуалната комуникация. Наблюдавайки лицето хората могат инстинктивно да извлекат много невербални послания, като например самоличността на човек, емоции и намерения. В компютърното зрение, за да се извлече тази информация от човешкото лице е необходимо да се локализират основните характерни точки на лицето (Фигура 1) и точността им е ключова за множество методи за лицев анализ. Например алгоритмите за разпознаване на изражението на лицето и ориентацията на главата разчитат много сериозно на информацията за формата на лицето в рамките на изображението предоставяна от тези характерни точки. Характерните точки, намиращи се около очите, могат да дадат първоначално предположение за позициите на центровете на зениците, които се използват за засичане и следене на погледа. За извършването на лицево разпознаване позициите на характерните точки върху двуизмерното изображение често се комбинират с триизмерен модел на главата за „изправяне" на лицето към фронтална позиция и по този начин да се намалят вариациите в рамките на външността на един индивид, което е ключово за повишаване точността на разпознаване [48]. Информацията, извлечена от лицето чрез тези характерни точки, е важна и при изграждането на интерфейс човек-машина, както и при видео охранителни системи и при медицински приложения.

Позициите на основните характерни точки на лицето около органите, намиращи се на лицето и неговите контури, отразяват евклидовите трансформации и неевклидовите деформации, причинени от движенията на главата и израженията на лицето. Поради това тяхната роля в различни задачи за разпознаване лица и анализ на поведение и емоции е изключително важна. Множество алгоритми са разработени за автоматични засичане на характерните точки на човешкото лице, като в тази дипломна работа правим обзор на най-значимите от тях.



***Фигура 1**: Примерни изображения с анотирани характерните точки на лицето. Източник: [48]*

Алгоритмите за засичане на характерни точки на лицето работят върху изображения или върху видеа, а точките които засичат са или доминантните точки на уникалните характеристики на органите намиращи се върху лицето – например ъгълчето на окото, или са интерполирани точки, свързващи тези доминантни точки около контурите на органите или на самото лице. Изказано формално, имайки дадено изображение на лице *I*, алгоритъм за засичане на характерни точки на лицето прогнозира местоположенията на *D* на брой характерни точки $x = \{x_1, y_1, x_2, y_2, ..., x_D, y_D\}$, където $x_.$ и $y_.$ представляват координати в рамките на даденото изображение.

Засичането на характерните точки на лицето все още е трудно предизвикателство по няколко причини. Първо, видът на лицето се изменя значително при различни позиции на главата и изражения. Второ, околните условия, като например осветеността, значително афектират вида на лицето върху снимки. Трето, закриването на лицето от други обекти като очила или маски, както и самозакриването в следствие на голямо извъртане на главата, водят до наличието на само частичен изглед на лицето върху снимките.

В последните години има значително развитие в алгоритмите за засичане на характерни точки на лицето и свързаните с това задачи. По-ранните разработки се фокусират върху по-малко предизвикателните случаи без гореизброените затрудняващи фактори – основно фронтални изображения с умерена осветеност. По-късните разработки се стремят да се справят с определени категории вариации и често използват изображения събрани в контролирани условия [48], например само определени позиции на главата и изражения на лицето. Най-новите разработки се фокусират върху предизвикателните „свободни" условия, при които разглежданите изображения могат да включват произволни изражения на лицето, позиции на главата, осветеност, закривания на части от лицето и т.н. Като цяло все още няма разработка, която да може постигне устойчивост на произволни условия от реалния свят.

Различни техники са били прилагани към задачата за засичане на характерните точки на лицето в зависимост с каква цел се решава тя. Например по-ранните разработки за засичане позата на главата и тялото на човек, преди напредъка на дълбоките невронни мрежи, са били основно базирани на принципа на напасване на деформируеми геометрични структури [16] и по-сложни негови модификации, поради възможността да се моделират по този начин големите изменения във външния вид причинени от разликите в ориентацията на главата. Този подход е показал голяма устойчивост на различните позиции на главата, но базираните на него разработки

не са достигнали по-високата точност, показана от разработките, използващи каскадна регресия [13], които пък от своя страна бързо губят точност при закриване на част от характерните точки в следствие на завъртане на главата.

С развитието на дълбокото машинно самообучение и създаването на големи анотирани набори от данни по-скорошните разработки, използващи конволюционни невронни мрежи (CNNs) постигат голям скок в точността при изображения, заснети в предизвикателните „свободни" условия. [4] прави обзор на лимитите в областта на засичането на характерни точки на лицето достигнати от съвременните архитектури на дълбоки невронни мрежи.

Скорошните разработки, използващи конволюционни невронни мрежи (CNNs), базирани на регресия на топлинни карти (heatmaps) [3, 44], успяват да постигнат висока точност дори на най-предизвикателните набори от данни, като благодарение на обучението на тези невронни мрежи от край до край без ръчна намеса, те лесно могат да бъдат още подобрени при създаването на нови, по-обширни набори от данни.

1.2 Цел и задачи на дипломната работа

Целта на настоящата дипломна работа е имплементация на система, която намира характерните точки на лицето върху статични цветни изображения, изрязани в малък регион около самото лице (правоъгълният регион резултат от предварителната работа на детектор за лица) и изчислява координатите им в пикселното пространство спрямо границите на този регион. Вторична цел е опростяване и оптимизиране на системата с цел постигане на работа в реално време на платформа за вградени системи. Това от своя страна поражда следните **задачи**:

1. Обзор на основните подходи за засичане на характерните точки на лицето.
2. Преглед на предходните разработки и на текущо достъпните набори от данни.
3. Предлагане на решение на тази задача, използващо дълбоки невронни мрежи.
4. Оптимизиране на системата с цел постигане скорост на обработка от поне поне 25 кадъра в секунда на платформа за вградени системи Nvidia Jetson TX2.
5. Планиране и провеждане на експерименти за оценка на предложеното решение спрямо стандартно приети метрики върху някои от разгледаните набори от данни.
6. Анализ на резултатите от експериментите.

1.3 Структура на дипломната работа

В глава 2 ще разгледаме по-подробно постановката на така въведената задача за засичане на характерни точки на лицето, както и основните приложения на резултатите от решението на тази задача. Ще направим обзор на трудностите в тази задача, представляващи предизвикателство пред текущите разработки на тази тематика. Ще представим обща характеристика на подходите за използвани за засичане на характерни точки на лицето в предходни разработки. В глава 3 ще разгледаме метриките използвани за оценка на точността на системите, решаващи тази задача, и ще направим обзор на достъпните набори от данни, които се явяват един от основните ограничаващи фактори при прилагането на методи за машинно самообучение за решаване на разглежданата задача. В глава 4 ще разгледаме значимите предходни разработки по темата и

постигнатите от тях резултати. В глава 5 ще представим избраните за решаване на поставените задачи алгоритми и архитектури на дълбоки невронни мрежи и разработената на тяхна база система за засичане, както и ще направи оценка на нейната работа . В глава 6 ще направим заключение и ще посочим насоки за бъдеща работа.

2. Преглед на проблемната област

2.1 Задача за засичане на характерни точки на лицето

Засичането на характерни точки на лицето, дефинирано като локализирането на координатите на определени ключови атрибути на лицето върху двуизмерно изображение е безспорно най-важната междинна стъпка за множество задачи при анализа на лица – от биометрично разпознаване до разчитане на емоции. Въпреки, че е концептуално проста, тази задача от областта на компютърното зрение се е показала като изключително сложна поради влиянието на редица фактори като поза на главата, изражение, закриване на части от лицето и осветеност.

Разглежданият проблем е следният – получавайки изображение на лице да се определят позициите на даден брой характерни точки съответстващи на най-важните характеристики на човешкото лице. Най-често използваните характерни точки са ъгълчетата на очите, върха на носа , ноздрите, краищата на устата, крайните точки на веждите, върха на брадата. Някои от тях като например ъгълчетата на очите и върха на носа е известно, че се влияят много слабо от изражението на лицето и затова се считат за по-надеждни и в специализираната литература се наричат fiducial points или fiducial landmarks [34]. Локализирането на тези точки е еквивалентно на разпознаването на основните органи намиращи се на лицето.

Характерните точки на лицето се разделят най-общо в два класа – основни точки, наричани още fiducial landmarks, и вторичните характерни. Основните точки, които се локализират директно, играят основната роля при лицевото разпознаване. Тези точки, например ъгълчетата на очите и устата, върха на носа и веждите могат да бъдат лесно засечени използвайки характеристики на изображението от по-ниско ниво. Втория клас – вторичните характерни точки са разположени по контурите свързващи основните точки – брадата, скулите, междинните точки по извивките на веждите и устните. Те играят важна роля при разпознаване на изражението и при следене движението на лицето.

Основна причина за нарастващия в последните години интерес към задачата за засичане на характерните точки на лицето е широкият спектър от актуални проблеми, за които тя е основна стъпка – от визуални ефекти, анимация на лица и триизмерна реконструкция на образи, през разчитане на изражения и жестове с глава до разпознаване на лица. Съществуващите комерсиални приложения използват характерните точки на лицето анонимизация на дигитални снимки и видеозаписи, четене по устните, подпомагане на автоматичния превод на езика на знаците [34] и други.

Главните четири сред задачите, зависими от точната локализация на характерните точки на лицето, са:

- Разчитане на изражения: израженията на лицето са визуален индикатор за емоциите и невербалните послания и играят важна роля като допълнение към вербалната комуникация. Пространственото разположение и динамиката на движението на характерните точки на лицето предоставят начин за обективно описване и анализ на жестовете с глава и израженията на лицето. На тази база автоматично могат да се разпознаят класове движения на лицето съответстващи на изразяваните емоции [2].
- Регистрация на лица: първоначалното регистриране на дадено лице е най-важният фактор оказващ влияние върху точността на последващото му разпознаване. Разглежданите характерни точки позволяват да се извлече надежден уникален идентификатор на дадено лице.
- Разпознаване на лица: системите за разпознаване на лице обикновено засичат позициите на очите и тогава извличат холистични характеристики от региони центрирани върху тях. Засечените характерни точки също така допринасят редица геометрични характеристики като разстояния и ъгъла между тях. Антропоморфните модели на лицето комбинират два източника на информация – геометричната конфигурация на характерните точки и визуалните характеристики около тях [34].
- Проследяване на лица: повечето алгоритми за проследяване на лицата във видеозаписи се подобряват включвайки проследяване на поредиците от характерни точки. Групата методи използваща напасване на графово представяне на лицето използва между 60 и 80 характерни точки. Проследяването след това се извършва чрез развитие на графа спрямо параметрично представяне на формата на лицето и на геометричните връзки между частите му. Алтернативният подход е да се изгради векторно поле описващо движенията на пикселите около засечените характерни точки [34].
- Други приложения на характерните точки на лицето включват триизмерна реконструкция на лицето от стереоизображения, множество снимки или видеозаписи, където те се използват за установяване на съответствие между точки в отделните кадри. Тези триизмерни модели позволят прилагането на специални ефекти като анимации, изменение на лицето, виртуален грим и други.

Реалната употреба на гореизброените приложения изисква алгоритмите за засичане на характерни точки на лицето да работят в реално време върху ограничената изчислителна мощ на вградени системи като например интелигентни камери за видеонаблюдение. Същевременно точността на засичане на характерните точки се отразява пряко на успеха на тези приложения.

Формално задачата за засичане на характерни точки на лицето може да бъде описана така: по дадено входно изображение I с размери $W \times H \times C$, където W е ширината, H – височината, C – броя на цветовите канали в изображението (обикновено 3, но може да бъде и 1 при

инфрачервени или черно-бели снимки), да се намери функция $\Phi: I \rightarrow L$, която от входното изображение I прогнозира вектор с характерните точки L, който за всяка характерна точка съдържа x и y координати в пространството на входното изображение. Броят на характерните точки може да е различен в зависимост от приложението, за което те се използват, и набора от данни използван за разработката на решението. Качеството на решението обикновено се определя от точността на функцията Φ върху тестов набор от данни [26].

2.2 Основни предизвикателства за засичане на характерни точки на лицето

Въпреки множеството разработки през последните няколко десетилетия, засичането на характерните точки на лицето остава много предизвикателен проблем. Основните предизвикателства, които ограничават точността и устойчивостта на алгоритмите, са следните:

- Локални вариации: изражението, локални екстремуми в осветеността (отблясъци и сенки) и закриване на части от лицето внасят частични изменения в заснетите изображения. В резултат регионите на някои от характерните точки може да са изместени от нормалното си местоположение или да са изчезнали от изображението.
- Глобални вариации: ориентацията на главата и качеството на снимките са двата основни фактора, които глобално афектират вида на лицата в заснетите изображения и могат да доведат до лоша точност при голяма част от засечените характерни точки, когато глобалната структура на лицето е грешно напасната към изображението.
- Все още няма разработен подход за засичане на характерните точки на лицето, който да е устойчив към глобални и локални вариации, и в същото време да има ефикасност, позволяваща изпълнение в реално време [34].
- Дисбаланс в данните: не е рядкост наборите от данни, използвани за създаването на модели чрез машинно самообучение, да имат дисбаланс в разпределението на данните между отделни класове (например фронтални срещу нефронтални снимки, светли срещу тъмни и т.н.). Този дисбаланс е много вероятно да попречи на обучавания модел да представи коректно характеристиките на данните, показвайки по-ниска точност при по-слабо представени данни.
- Условията на заснемане на изображението: фактори като например осветеност, резолюция и детайли на фона могат да повлияят на точността на засичане на характерните точки на лицето. Това ясно се вижда от факта, че често модели обучени върху един набор от данни, дават по-лош резултат, когато са тествани върху друг набор от данни. [12] изследва задълбочено влиянието върху точността на засичане от няколко фактора като резолюция, изражение, частично закриване на лицето, избран модел и набор от данни използван за обучението му.
- С високото разпространение на мобилните устройства и интернет на нещата, все повече задачи се изпълняват върху преносими или вградени устройства. Затова важна цел пред алгоритмите за засичане на характерни точки на лицето, освен високата точност, е и ефикасността на изпълнението им, и в частност ефикасност при изпълнение върху платформи с ограничена изчислителна мощ и памет, като например смартфони, което

изисква използваните модели за машинно самообучение да бъдат с по-малки размери, сложност и брой параметри.

- Броят на засичаните характерни точки на лицето и изискванията за тяхната точност зависят от търсеното им приложение. Основните характерни точки като тези около очите и носа е необходимо да бъдат по-точно локализирани, тъй като те често се използват за насочване на локализацията на вторичните, по-слабо и ненадежно изразени върху изображението, точки. Наблюдава се обаче, че характерните точки по контура на лицето, като например върху брадичката не могат да бъдат анотирани точно, независимо ръчно или автоматизирано. В резултат в специализираната литература се разглеждат като основни 17 характерни точки, разположени навътре от контура на лицето – 4 по веждите, 6 по очите, 3 на носа и 4 на устата. Тези точки често се групират заедно и означават като m17 в литературата [34].

[48, 26] правят подробни оценки и сравнения на по-известните предишни разработки в областта с оглед степента на разрешаване на изложените по-горе предизвикателства. [48] се фокусира върху групиране на разработките спрямо използвания подход в 3 главни категории, които разглеждаме подробно в глава 2.3, а [26] се съсредоточава върху използваните функции на грешката (loss functions).

[4, 26, 34] разглеждат публично достъпните набори от данни и правят оценка доколко вариацията на данните в тези набори ограничава справянето с част от разгледаните предизвикателства при използване на методи базирани на машинно самообучение и по-конкретно дълбоко самообучение.

[21] предлага архитектура на дълбока невронна мрежа за засичане на характерни точки на лицето специално предвидена за работа в реално време върху мобилни и вградени устройства и прави сравнителна оценка на множество предходни разработки в контекста не само на точност, но и на приложимост и ефикасност при изпълнение върху мобилни устройства.

2.3 Подходи за засичане на характерни точки на лицето

Алгоритмите за засичане на характерни точки на лицето могат да бъдат разделени на три основни категории в зависимост от това как те моделират външния вид и формата на лицето: холистични методи, ограничени локални методи (CLM – Constrained Local Model) и методи базирани на регресия. В този контекст под външния вид на лицето се има предвид специфичните модели, формирани от цвета и интензитета на пикселите около характерните точки, както и по лицето като цяло, докато под форма се разбира геометричния модел на лицето, формиран от позициите на характерните точки и тяхното взаимно разположение в пространството. Холистичните методи експлицитно моделират цялостния външен вид на лицето и цялостната му форма с глобален модел на геометричната структура формирана от характерните точки. Ограничените локални методи разчитат на експлицитно моделиране на външния вид само на локални региони около характерните точки, но също формират модел на цялостната геометрична структура формирана от характерните точки. Методите основани на регресия може да използват и локален, и глобален модел на външния вид на лицето, а формата може да е отразена имплицитно

чрез едновременното локализиране на всички характерни точки. Таблица 1 обобщава разликите между трите класа алгоритми и основните характеристики на базираните на тях разработки:

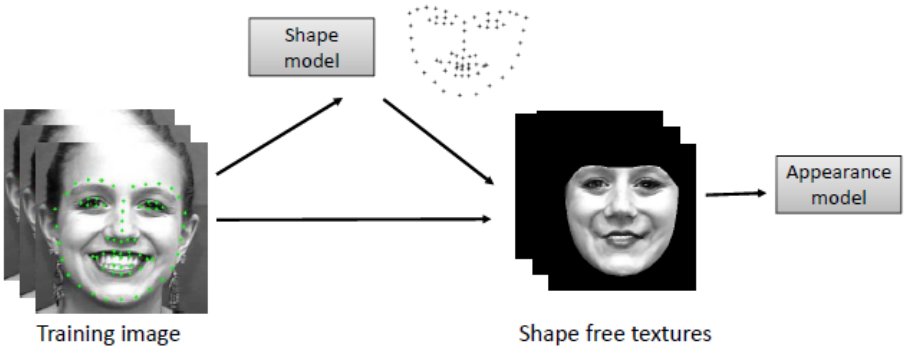
Клас алгоритми	Външен вид	Форма	Точност и устойчивост	Скорост
Холистични методи	цялото лице	експлицитно	лоша генерализация/добра	бавни/бързи
Ограничени локални методи (CLM)	локални региони	експлицитно	добра	бавни/бързи
Методи основани на регресия	локални региони и/или цялото лице	имплицитно	добра/много добра	бързи/много бързи

***Таблица 1**: Сравнение на основните класове алгоритми за засичане на характерни точки на лицето.*
Източник: [48]

Най-добра точност се постига от разработките базирани на регресия [48], каквито са много от по-съвременните разработки. Някои скорошни разработки комбинират дълбоко машинно самообучение с глобален триизмерен модел на формата на лицето и по този начин излизат извън обхвата на тези категории. Тях ще разгледаме отделно.

2.3.1 Холистични методи

Холистичните методи използват информация за цялостният външен вид на лицето, както и глобалната му структура за засичане на характерните му точки (Фигура 2). Първо ще разгледаме една от най-ранните разработки от тази категория - считания за класически холистичен метод Active Appearance Model (AAM) [9]. Той служи за основа на множество доработки, които ще разгледаме след него.



***Фигура 2**: Принцип на работата на холистичните методи.*
Източник: [48]

Методът **Active Appearance Model (AAM)** е представен от Taylor и Cootes [9, 15]. Това е статистически метод, който напасва към изображенията на лица малък брой коефициенти отразяващи и външния вид и вариациите във формата. По време на конструиране на модела AAM изгражда последователно модели на глобалната форма на лицето и на цялостния му изглед използвайки анализ на основните компоненти (Principal Component Analysis, PCA). По време на засичане на характерните точки алгоритъма напасва научените модели към тестовите изображения.

Няколко са стъпките, чрез които AAM конструира моделите на външния вид и геометричната форма на лицето по подадени обучителни изображения с анотирани характерните точки на лицето, означени като $\{I_i, x_i\}_{i=1}^{N}$, където N е броя на обучителните изображения. Първо се прилага Procrustes Analysis [1], за да се регистрират формите на всички лица от обучителните. Премахват се афинните трансформации на формата на всяко лице x_i и се генерират нормализирани обучителни форми x_i'. След това, имайки нормализираните обучителни форми $\{ x_i'\}_{i=1}^{N}$, се прилага анализ на основните компоненти (PCA), за да се научи средната форма s_0и ортонормиран базис $\{ s_n\}_{n=1}^{K_s}$, където K_s е размерността на използвания базис (фигура 3). Имайки научения базис за представяне на формата на лицето базис $\{ s_n\}_{n=0}^{K_s}$, една нормализирана форма x' може да бъде представена чрез коефициентите $p = \{ p_n\}_{n=1}^{K_s}$ по следния начин:

$$x' = s_0 + \sum_{n=1}^{K_s} p_n * s_n \quad (1)$$

***Фигура 3**: Научени вариации във формата чрез AAM.*
Източник: [21]

***Фигура 4**:Научени вариации във външния вид чрез AAM.*
Източник: [21]

След това, за да се научи модела на външния вид, изображенията се деформират, за да се напаснат към осреднената форма и да се генерират нормализирани по формата изображения означени като $\mathrm{I}_\mathrm{i}\big(W(x_i')\big)$, където $W(.)$ означава операцията за деформация на изображението. Тогава отново се прилага анализ на основните компоненти (PCA) на нормализираните изображения на лицата $\big\{\mathrm{I}_\mathrm{i}\big(W(x_i')\big)\big\}_{i=1}^{N}$, за да се научат средният външен вид A_0 и базис с размерност K_a за представяне на външния вид $A = \{ A_m\}_{m=1}^{K_a}$, както е показано на фигура 4. Имайки модела на външния вид $A = \{ A_m\}_{m=0}^{K_a}$, всяко нормализирано по формата изображение може да бъде представено чрез коефициентите $\lambda = \{ \lambda_m\}_{m=1}^{K_a}$ чрез формулата:

$$\mathrm{I}\big(W(\mathrm{x}')\big) = A_0 + \sum_{m=1}^{K_a} \lambda_m * A_m \quad (2)$$

Опционално, трети модел може да се приложи, за да се научи връзката между коефициентите на формата *p* и на външния вид λ [48].

При засичане на характерните точки на лицето AAM намира коефициентите на формата и на външния вид p и λ, параметрите на афинната трансофрмация (c, θ, t_c, t_r означават параметрите на мащабиране, ротация и транслация), които най-добре напасват подаденото изображение и те определят локациите на характерните точки:

$$x = cR_{2d}(\theta)(s_0 + \sum_{n=1}^{K_s} p_n * s_n) + t \qquad (3)$$

Тук $R_{2d}(\theta)$ означава матрицата на ротация, а $t = \{t_c, t_r\}$. За да се опрости представянето по-надолу в текста коефициентите на формата ще включват и коефициентите установени от PCA анализа и параметрите на афинната трансформация.

Обобщено, процедурата по напасване може да се формулира като минимизиране на разстоянието между реконструираните изображения $A_0 + \sum_{m=1}^{K_a} \lambda_m * A_m$ и нормализираното по формата входно изображение $\mathrm{I}(W(p))$. Разликата между двете обикновено се нарича изображение на грешката, означено като ΔA:

$$\Delta A(\lambda, p) = Diff\left(A_0 + \sum_{m=1}^{K_a} \lambda_m * A_m, \mathrm{I}(W(p))\right) \qquad (4)$$

$$\lambda^*, p^* = \arg\min_{\lambda,p} \Delta A(\lambda, p) \qquad (5)$$

При класическият AAM [9, 15], коефициентите на модела се намират чрез итеративно пресмятане на изображението на грешката въз основа на текущите коефициенти и преизчисляване на прогнозираното им обновление на база на изображението на грешката.

Повечето от холистичните методи, основани на AAM се фокусират върху алгоритмите за напасване на изображенията към научените модели, което включва решаването на уравнение 5. Те могат да бъдат разделени на аналитични методи за напасване и методи, основани на самообучение.

Аналитичните методи за напасване формулират задачата за напасване на изображенията от AAM като нелинеен оптимизационен проблем и я решават аналитично. По конкретно, алгоритъмът търси най-добрия набор от коефициенти на формата и външния вид p и λ, които минимизират разликата между реконструираното изображение и входното изображение с нелинейна формулировка по метода на най-малките квадрати:

$$\tilde{\lambda}, \tilde{p} = \arg\min_{\lambda,p} \left\| A_0 + \sum_{m=1}^{K_a} \lambda_m * A_m - \mathrm{I}(W(p)) \right\|_2^2 \qquad (6)$$

Тук, $A_0 + \sum_{m=1}^{K_a} \lambda_m * A_m$ представлява реконструираното лице, нормализирано по формата, спрямо параметрите на формата и на външния вид, а цялата целева функция представлява грешката при реконструкцията.

Вместо директно да решат аналитично задачата за напасване на изображенията, **методите базирани на самообучение** научават как да прогнозират коефициентите на формата и външния вид от вида на лицето в изображенията. Те могат допълнително да бъдат разделени на методи за напасване чрез линейна регресия, чрез нелинейна регресия и други методи.

Методи за напасване чрез линейна регресия: тази група методи приемат, че съществува линейна зависимост между обновленията на коефициентите на модела и изображението на грешката $\Delta A(\lambda, p)$ или характеристиките на самото изображение $I(\lambda, p)$. Те научават функцията на линейна регресия от предвижданията на модела следвайки класическия AAM алгоритъм описан по-горе.

$$\Delta A(\lambda, p)\ or\ \Delta I(\lambda, p) \xrightarrow{linear\ regression} \Delta\lambda, \Delta p \qquad (8)$$

Така те определят коефициентите на модела като итеративно изчисляват обновленията в тези коефициенти и ги добавят към текущите стойности, след което извършват предвиждане за да започнат следващата итерация.

Методи за напасване чрез нелинейна регресия: методите основани на линейна регреси допускат, че зависимостта между характеристиките на изображението и изображението на грешката, описана в уравнение 4, около истинското решение за коефициентите на модела е квадратична, което означава, че итеративна процедура с линейни обновления и адаптивен размер на стъпката ще доведе до сходимост. Но това допускане за линейност е вярно само когато начална инициализация е близка до истинското решение, което прави методите за напасване основани на линейна регресия много чувствителни към инициализацията на коефициентите. За да се справят с този проблем, методите, основани на нелинейна регресия, използват нелинейни функции за моделиране на връзката между характеристиките на изображението и обновленията на коефициентите на модела:

$$\Delta A(\lambda, p)\ or\ \Delta I(\lambda, p) \xrightarrow{nonlinear\ regression} \Delta\lambda, \Delta p \qquad (9)$$

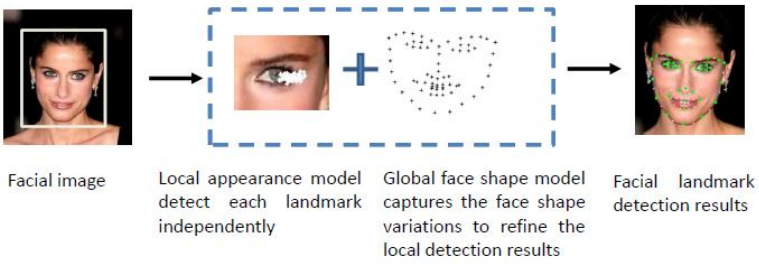
[45] сравнява версии на AAM с линейни и нелинейни алгоритми за регресия. Авторите емпирично показват, че нелинейните методи са по-добри в първите няколко итерации за избягване на локални минимуми, докато линейните методи са по-добри, когато приближението е близо до истинското решение.

Други методи за напасване базирани на самообучение: съществуват други доработки на оригиналния AAM алгоритъм. Една конкретна посока на развитие е да се подобри представянето на характеристиките на изображението. Известно е, че AAM базираните модели имат ограничена възможност за генерализиране и трудно напасват лица с вариации несрещащи се при обучителните данни (например осветеност, частично закриване и т.н.) [19, 20]. Това ограничение е отчасти поради директното използване на интензитета на пикселите като характеристики на изображението. За да се справят с този проблем, някои алгоритми използват по устойчиви на вариации характеристики, например в [23] вместо директно интензитета на пикселите използват wavelet функции за моделиране на външния вид на лицата.

Ансамбли от AAM модели: единичен AAM модел неизбежно предполага линейност във вариациите на формата и външния вид на лицата. Поради това ограничение методи използват ансамбъл от модели за да подобрят устойчивостта и точността. Например в [40] е предложен AAM модел с последователна регресия, който обучава серия от AAM модели за последователно напасване в каскаден стил. Моделите в по-ранните етапи отчитат големите вариации, например ориентацията на главата, докато тези в по-късните етапи компенсират малките вариации. В тази публикация се разглеждат и ансамбли от самостоятелни AAM модели, и от последователно свързани такива.

2.3.2 Ограничени локални методи (CLM)

Както е показано на фигура 5, ограничените локални методи извеждат позициите на характерните точки *x* въз основа на глобалните геометрични модели във формата на лицето, както и на независима информация за локалния външен вид на лицето около всяка характерна точка [10, 39], който е по-лесен за моделиране, както и по-устойчив на осветеност и закриване в сравнение с цялостния външен вид на лицето.



***Фигура 5**: Принцип на работата на ограничените локални методи.*
Източник: [48]

Погледнато генерално, ограничените локални методи могат да бъдат формулирани или като детерминистични, или като вероятностни методи. При детерминистичната гледна точка, методите намират характерните точки на лицето чрез минимизиране грешката на напасване в геометричните структури:

$$\tilde{\boldsymbol{x}} = \arg\ \min_x Q(x) + \sum_{d=1}^{D} D_d(x_d, I) \qquad (10)$$

Тук x е вектора с характерните точки, а x_d означава позициите на отделните точки от от него. $D_d(x_d, I)$ представлява локалната оценка на доверието в x_d . $Q(x)$ представлява регуляризация наказваща невъзможни или неантропологични форми на лицето в контекста на глобалната му структура. Интуицията зад тази формула е, че искаме да намерим най-добрия набор от характерни точки на лицето, който максимизира независимото локално доверие в позицията на всяка точка и удовлетворява ограничението на глобалната геометрична структура на човешкото лице.

Регуларизацията на формата може да бъде приложена към коефициентите на формата в модела **p**, ако я означим с $Q_{\mathbf{p}}(\mathbf{p})$ уравнение 10 се записва като:

$$\tilde{\mathbf{p}} = \arg\ \min_{\mathbf{p}} Q_{\mathbf{p}}(\mathbf{p}) + \sum_{d=1}^{D} D_d(x_d(\mathbf{p}), I) \qquad (11)$$

При вероятностната гледна точка, ограничените локални методи могат да се разглеждат кат максимизиране на произведението на априорната вероятност на геометричната форма на лицето $p(x;\eta)$, състояща се от всички характерни точки и на вероятността на локалния външен вид на лицето около всяка точка $p(x_d|I;\theta_d)$:

$$\tilde{x} = \arg\ \max_x p(x;\eta) \prod_{d=1}^{D} p(x_d|I;\theta_d) \qquad (12)$$

Подобно на детерминистичната формулировка, априорната вероятност може да бъде приложена към коефициентите на формата в модела **p** и уравнение 12 се записва като:

$$\tilde{\mathbf{p}} = \arg\ max_{\mathbf{p}}\, p(\mathbf{p};\eta) \prod_{d=1}^{D} p(x_d(\mathbf{p})|I;\theta_d) \qquad (13)$$

И при детерминистичните и при вероятностните ограничени локални методи има два основни компонента. Първият компонент е моделът на локалния външен вид отразен в $D_d(x_d, I)$ или $p(x_d|I;\theta_d)$, в уравнения 10, 11, 12 и 13 съответно. Втората компонента отразява

ограниченията върху геометричната структура на формата на лицето, приложена или към коефициентите на формата в модела **p**, или към самата форма представлявана от вектора с координати на характерните точки **x**, като регуларизационен член или като априорно вероятностно разпределение. Двата компонента обикновено се научават по отделно във фазата на обучение на модела и се комбинират при извеждане на координатите при засичане характерните точки.

Моделът на локалния външен вид пресмята коефициент на доверие $D_d(x_d, I)$ или вероятност $p(x_d|I;\theta_d)$, че характерната точка с индекс d се намира на конкретните пикселни координати x_d въз основа на локалния външен вид около x_d в изображението I. Моделите на локалния външен вид могат да бъдат разделени на модели основани на класификация и модели основани на регресия.

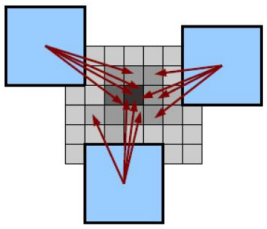
Моделите, основани на класификация на локалния външен вид, обучават бинарен класификатор да разпознава положителни парчета от изображението, центрирани в истинската позиция на съответните характерни точки от отрицателни парчета, които са отдалечени от анотираната позиция. По време на засичане на характерните точки класификаторът може да се прилага към различни позиции в изображението за да генерира коефициенти на доверие $D_d(x_d, I)$ или вероятности $p(x_d|I;\theta_d)$ чрез гласуване. Използват се различни характеристики на изображението и различни класификатори.

По време на обучение, целта на основаните на регресия модели на локалния външен вид е да предвидят вектор на отместване $\Delta \mathrm{x}_\mathrm{d}^* = x_\mathrm{d}^* - \mathrm{x}$, който представлява разликата между произволни пикселни координати x и анотираните истински координати на характерната точка x_d^*, използвайки локалната информация за външния вид около x, чрез използване на регресия. По време на засичане на характерните точки регресорът може да се прилага към различни позиции в даден регион от изображението за да предвиди $\Delta \mathrm{x}_\mathrm{d}$, което да бъде добавено към текущата позиция x за да се изчисли x_d.

$$Regression: I(x) \rightarrow \Delta x_\mathrm{d} \qquad (14)$$

$$x_\mathrm{d} = \mathrm{x} + \Delta x_\mathrm{d} \qquad (15)$$

Предвижданията от множество региони могат да се комбинират, за да се получи крайното предвиждане за коефициента на доверие или вероятността чрез гласуване.



***Фигура 6**: Принцип на работата на модел на локалният външен вид основан на регресия. Източник: [48]*

Моделът на формата на лицето описва пространствените връзки между характерните точки на лицето, които ограничават и насочват търсенето на координатите им. Най-общо тези модели могат да се разделят на детерминистични и вероятностни модели на формата.

Детерминистичните модели на формата използват детерминистично описание на геометричните структури във формата на лицето. Те слагат ниска стойност за грешката при напасване на реалистични форми на лицето и висока на непостижими форми. Например алгоритъмът Active shape model (ASM) е един от най-популярните класически модели на формата на лицето [8]. Той научава линейните подпространства на формите на лицата чрез анализ на основните компоненти (PCA) както в уравнение 1. Оценява се напасването на лицето към тези подпространства. Този подход се използва както в холистични метод AAM, така и в някои ограничени локални методи. Тъй като един линеен ASM модел може да не отрази ефективно глобалните вариации във формата на лицето, в [31] конструират две нива на ASM модели. Едното ниво отразява геометричните структури на всеки компонент от лицето по отделно, а другото моделира тяхната обща пространствена връзка.

Вероятностните модели на формата на лицето отразяват геометричните структури във формата чрез назначаване на високи вероятностни на форми, удовлетворяващи антропологичните ограничения научени от обучителните данни, и ниски вероятности на непостижимите форми. Една от ранните вероятностни разработки [43] използва превключване между дискретни състояния за отделните части на лицето, за да се справи с различни изражения. Например автоматично преминава от състояние на отворена уста към състояние на затворена според постъпващите данни.

2.3.3 Методи основани на регресия

Методите основани на регресия директно научават съпоставяне от външния вид на изображението към координати на характерните точки на лицето. За разлика от холистичните и ограничените локални методи, те обикновено не изграждат глобален модел на формата на лицето. Вместо това ограниченията за формата на лицето се включват имплицитно в модела. Най-общо методите основани на регресия могат да се разделят на ползващи директна регресия, ползващи каскадна регресия и основани на дълбоко самообучение. Ползващите директна регресия предвиждат координатите на характерните точки с една итерация и без инициализация, докато каскадните регресори извършват предвижданото последователно в каскада и обикновено

имат нужда от инициализация за позициите на точките. Методите основани на дълбоко самообучение използват и двата подхода и затова ще ги разгледаме отделно.

2.3.3.1 Методи с директна регресия

Методите основани на директна регресия научават директно съпоставяне от вида на изображението към координатите на характерните точки без никаква инициализация за координатите. Обикновено се изпълняват в една единствена стъпка. Могат допълнително да се разделят на локални подходи и глобални подходи. Локалните подходи използват части от изображението, докато глобалните подходи използват цялостния външен вид на лицето.

Локални подходи: локалните подходи семплират различни части от региона на лицето и създават структурирани регресори за предвиждане на вектори на отместването (между целевата форма на лицето и извлечените региони), които могат да се добавят към текущите позиции на регионите за съвместно изчисление на всички характерни точки. Крайните координати на характерните точки на лицето могат да се изчислят чрез комбиниране на предвижданият от множество семплирани региони. Трябва да се отбележи, че тези подходи се отличават от основаните на регресия модели на локалния външен вид разгледани по-горе в секция 2.3.2, които прогнозират всяка характерна точка независимо от останалите, докато разглежданите тук локални подходи прогнозират координатите на всички точки съвместно. Например в [11] авторите използват гора от условни регресори за да научат съответствието от случайно семплирани части от региона на лицето към обновленията на формата на лицето. В допълнение изграждат няколко отделни модела в зависимост от ориентацията на главата и ги комбинират заедно за засичане на характерните точки на лицето. Подобно в [51] използват гора от условни регресори въз основа на „привилегирована информация" – данни достъпни само по време на обучение на модела – в случая допълнителни характеристики на лицето (пол, позиция на главата и други). За разлика от метода в [11], които комбинира прогнозите на отделните модели за различни позиции на главата, при засичане този модел първо прогнозира допълнителните атрибути, а след това на тяхна база извършва локализирането на характерните точки. В този случай точността на засичане на координатите е пряко повлияна от точността на предвиждане на атрибутите. Един основен проблем на локалните подходи е, че независимите отделни региони може да не предоставят достатъчно информация за глобалната структура на формата на лицето. В допълнение, при закриване на части от лицето, случайно семплираните региони могат да доведат до грешни предвиждания.

Глобални подходи: глобалните подходи научават директно съответствие между цялостното изображение на лицето и координатите на характерните точки. За разлика от локалните подходи, цялостното лице предоставя повече информация за характерните точки, но научаването на съответствие между глобалния вид на лицето и позициите на характерните точки е по-трудно, тъй като цялостният външен вид на лицето има много повече вариации и е по-податлив на частично закриване. Всички водещи подходи от тази група използват дълбоко самообучение за научаване на съответствието [48], което ще разгледаме в секция 2.3.3.3. Отбелязваме, че тъй като глобалните подходи директно прогнозират позициите на характерните

точки, те се различават от холистичните подходи разгледани в секция 2.3.1, които конструират модели за формата и за външния вид и прогнозират коефициентите в тези модели.

2.3.3.2 Подходи използващи каскадна регресия

За разлика от методите основани на директна регресия, които правят прогнозиране в една стъпка, каскадните регресори започват с начално предположение за позициите на характерните точки (например осреднено представяне на лицето) и постепенно обновяват позициите на точките през фази с различни функции за регресия научени за различните етапи на предвиждане (фигура 7). Специфично при тях е, че при обучение на модела на всеки етап модели за регресия се прилагат за научаване на съответствие между семплиран според формата външен вид (локалния външен вид на лицето извлечен от текущото предположение за позициите на характерните точки) и обновлението в предположенията за позициите на характерните точки. Моделите, научени при по-ранните етапи, се използват за обновяване на обучителните данни за следващите етапи. По време на работа научените регресори се прилагат последователно за обновяване на предвижданото на всяка итерация.

***Фигура 7**: Принцип на работата на методите ползващи каскадна регресия.*
Източник: [48]

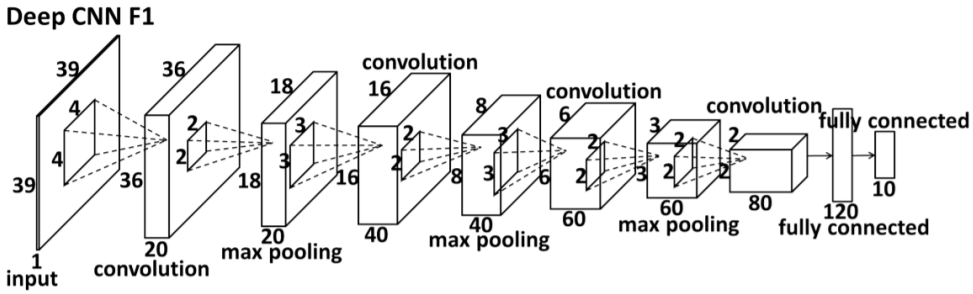
В тази категория има множество разработки, включващи различни видове регресия, както и различни подходи за семплиране на външния вид спрямо текущото предположение за позициите на характерните точки.

2.3.3.3 Методи основани на дълбоко самообучение

В последно време методите на дълбокото самообучение станаха един от най-популярните инструменти за решаване на задачи на компютърното зрение. Най-съвременните разработки в областта на засичането на характерните точки на лицето преминават от класическите подходи, които разгледахме по-горе към методи основани на дълбоко самообучение. В една по-ранните такива разработки - [49] използва дълбок вероятностен модел – машина на Болцман за отразяване на вариациите във формата на лицето, причинени от изражението му и позата на главата. В по-скорошните разработки моделите, използващи конволюционни невронни мрежи (CNNs), доминират в областта на засичането на характерните точки на лицето и повечето от тях

следват рамките на глобалните подходи с директна регресия или на тези с каскадна регресия. Обобщено, те могат да се разделят на чисто самообучителни методи и на хибридни методи. Чисто самообучителните методи предвиждат директно координатите на характерните точки, докато хибридните методи комбинират дълбокото самообучение с модели на проекция, широко използвани при компютърното зрение [48].

Чисто самообучителни методи: методите в тази категория използват конволюционни невронни мрежи, за да предвидят директно позициите на характерните точки от входните изображения. [41] е сред най-ранните разработки от тази група и предвижда 5 основни характерни точки на лицето в каскаден стил. Първото ниво включва CNN моделс 4 конволюционни слоя (фигура 8) за предсказване на позициите от входното изображение, след което няколко по-плитки мрежи локално прецизират всяка от отделните характерни точки.



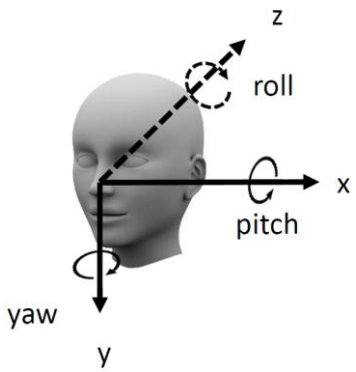
***Фигура 8**: Структура на CNN модела.*
Източник: [41]

Впоследствие други разработки подобряват [41] в две посоки. В едната използват многоцелево самообучение за подобрение на точността. Идеята е, че множество задачи може да имат единно представяне и съвместната им връзка може да подобри точността на отделните задачи. Няколко разработки комбинират многоцелево самообучение с CNN модели за съвместно предвиждане на характерните точки, позата на главата и атрибути на лицето. Други комбинират предвиждането на характерните точки с намирането на самото лице в по-голямо изображение и пол, като използват отделни комволюции за по-големи и по-малки характеристики на изображението [48].

Във втората посока, някои разработки подобряват каскадната процедура на [41]. Една разработка използва много сходен модел, за да предвиди 68 характерни точки на лицето вместо само 5. Друга използва дълбок автоенкодер за същото каскадно локализиране на характерните точки. Трета имитира каскадния подход, използвайки една единствена рекурентна невронна мрежа за локализирането на координатите от начало до край. Каскадните етапи са вградени в различните времеви отрязъци от рекурентната мрежа [48].

Хибридни методи: хибридните методи с дълбоко самообучение комбинират CNN моделите с триизмерни модели като проекционен модел или 3D деформируем модел на формата на главата (фигура 9). Вместо директно да предвиждат 2D координатите на характерните точки, те

предвиждат коефициентите на 3D деформируем модел на формата на лицето и ориентацията на главата. След това 2D координатите могат да бъдат изведени чрез модел на проекция характерен за компютърното зрение. Например в [53] изграждат плътен 3D модел на формата на лицето, след което итеративна каскадна регресия и дълбок CNN модел се използват за обновяване на коефициентите на 3D формата на лицето и ориентацията на главата. При всяка итерация, за да се отчете текущото приближение на 3D параметрите, 3D формата се проектира в 2D, и 2D формата на лицето се използва като допълнителен вход за CNN модела за предвиждането чрез регресия.



***Фигура 9**: 3D модел на лицето и неговата проекция въз основа параметрите на ориентацията на главата (ъгли pitch, yaw, roll).*
Източник: [48]

Сравнени с чисто самообучителните методи, хибридните методи с 3D деформируем модел и параметри на ориентацията на главата са по-компактен начин за представяне на двуизмерните координати на характерните точки на лицето. Поради това има по-малко параметри за изчисление и ограниченията за формата на лицето могат да бъдат експлицитно отразени. Освен това поради наличието на параметри за ориентацията на главата в 3D, те се справят по-добре с вариациите в ориентацията [48].

CNN моделите за засичане на характерни точки на лицето обикновено съдържат 4 конволюционни слоя и един напълно свързан слой. Сложността на моделите е сходна с дълбоките модели, използвани за други задачи за анализ на лица, като например – определяне ориентацията на главата, определяне на възраст и пол, разпознаване на изражения на лицето – които обикновено имат сходен или по-малък брой конволюционни слоеве [48]. За задачите за разпознаване на лица, CNN моделите обикновено са по-сложни, с повече конволюционни слоеве и напълно свързани слоеве, като това е отчасти поради наличието на много повече обучителни данни (в порядъка на десетки милиони изображения) в наборите от данни за разпознаване на лица в сравнение с тези за засичане на характерни точки на лицето (достигащи до около 20000 изображения) [48]. Все още е отворен въпросът дали добавянето на повече данни ще подобри работата на методите за засичане на характерни точки на лицето. До някъде този въпрос е разгледан в [4]. Друга обещаваща посока за работа е използването на идеята за многоцелево самообучение за съвместно предвиждане на свързани задачи (например ориентация на главата, възраст и пол) с по-дълбок модел, което да подобри точността на всички задачи.

2.3.3.4 Обобщение на методите основани на регресия

Сред различните методи основани на регресия, методите използващи каскадна регресия постигат по-добри резултати от методите, използващи директна регресия. Каскадните регресори, използващи дълбоко самообучение, постигат допълнително подобрение. Един проблем за всички методи използващи регресия е, че са чувствителни към детектора на лица, който е използван за определяне на ограждащия лицето регион, тъй като те научават съответствие между външния вид на лицето в рамките на този регион и позициите на характерните точки на лицето. Тъй като началната инициализация за позициите на точките е определена от ограждащия регион, модели обучени с един детектор може да не работят добре с друг детектор, имащ различно отклонение (bias).

Въпреки че както казахме, методите основани на регресия не изграждат експлицитно модел на геометричната форма на лицето, тази форма обикновено е имплицитно отразена от модела. На практика, тъй като методите, използващи регресия, прогнозират всички характерни точки съвместно структурната информация и ограниченията за формата на лицето се научават имплицитно в този процес.

2.4 Свързаност между трите основни категории подходи

В предишните няколко секции разгледахме алгоритмите за засичане на характерните точки на лицето разделени в три основни категории – холистични методи, ограничени локални методи и методи основани на регресия. Съществуват прилики и връзки между тези три категории.

Първо – и холистичните, и ограничените локални методи отразяват глобалните структури в геометричната форма на лицето чрез експлицитно създадени модели за формата, като подходите за това често са споделени между двете категории. Ограничените локални методи надграждат над холистичните методи като използват локалния външен вид около характерните точки вместо цялостния вид на лицето. Причината за това е, че е по-трудно да се моделира цялостния вид на лицето и че локалните региони от изображението са по-устойчиви на промени в осветеността и закриване на части от лицето в сравнение с цялостните модели използвани от холистичните методи.

Второ - методите основани на регресия, особено тези използващи каскадна регресия, се основават на същата интуиция като холистичния метод AAM, който разгледахме в началото. Например и в двата случая позициите на характерните точки на лицето се извеждат чрез напасване на външния вид на лицето и изчисленията могат да бъдат формулирани като нелинейна оптимизация по метода на най-малките квадрати, както беше показано в уравнение 6. Но холистичните методи предвиждат коефициентите на моделите на 2D формата и външния вид на лицето чрез напасване на цялостния модел на външния вид, докато каскадните регресори предвиждат директно характерните точки на лицето чрез напасване на локални модели на външния вид без експлицитно моделиране на 2D формата на лицето. Задачата за напасване при холистичните методи може да бъде решена аналитично или чрез самообучение, както разгледахме в края на секция 2.3.1, докато всички регресори се основават на самообучение. Докато основаните на самообучение методи за напасване при холистичните модели обикновено

използват един и същи модел за обновление на коефициентите в итеративен стил, каскадните регресори научават различни модели в каскаден стил. Алгоритъмът AAM [9] разгледан в началото на секция 2.3.1 е един конкретен тип холистичен модел, който е много подобен model на метода Supervised Descent Methods (SDM) [50] – един конкретен тип каскаден регресор. И двата обучават каскадни модели за научаване на съответствието между семплирани по текущото приближение на формата характеристики на изображението и обновлението в коефициентите на модела на формата на лицето. Обученият модел в текущия етап на каскадата ще модифицира обучителните данни преди обучението на модела от следващия етап на каскадата. Докато холистичния метод напасва цялостния външен вид на лицето и предвижда коефициентите на модела, SDM напасва локалния външен вид и предвижда директно позициите на характерните точки.

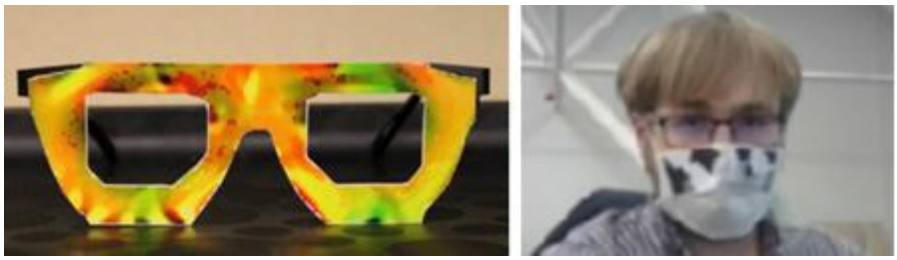
Трето – съществуват прилики между основаните на регресия модели на локалния външен вид използвани при ограничените локални методи (CLM), разгледани в секция 2.3.2 и методите основани на регресия разгледани в секция 2.3.3. И в двата случая обновленията в позициите на точките се предвиждат от първоначално предположение за позициите им. Първият подход предвижда всяка характерна точка независимо от останалите, докато вторият ги предвижда съвместно, така че ограниченията върху формата на лицето да бъдат имплицитно отразени. Първият тип подходи обикновено използват предвиждане в една стъпка с един регресор, докато вторият тип могат да използват различни регресори в каскаден стил.

Четвърто – сравнени с холистичните мтоди и ограничените локални методи, методите основани на регресия са по-обещаващи. Те заобикалят експлицитното моделиране на геометричните структури на формата на лицето и отразяват антропологичните ограничения върху формата имплицитно. Основаните на регресия методи директно предвиждат характерните точки, вместо да предвиждат коефициенти на модела, както правят холистичните и някои от ограничените локални методи. Директното предвиждане на позициите често е по-точно, тъй като малки изчислителни грешки в коефициентите на модела могат да доведат до големи грешки в крайните позиции на характерните точки.

2.5 Неразрешени проблеми при съществуващите решения за засичане на характерни точки на лицето

Съвременните алгоритми за компютърно зрение, включително използващите невронни мрежи, са податливи на така наречените „противникови атаки” (adversarial attacks), за пръв път разгледани по отношение на задачи на компютърното зрение в [42], където чрез добавяне на специално изработен шум към дигитални изображения (незабележим за човешкото око), авторите успяват драстично да променят предвижданията на невронна мрежа в задача за класификация на изображения. Тази атака е постигната чрез максимизиране на грешката на мрежата върху тестовото изображение L-BFGS метод. При тестване на мрежата върху такъв „противников пример”, генериран за набора от данни MNIST, те са успели да накарат мрежата да класифицира грешно почти всички примери. Трябва да се подчертае, че при такъв тип атаки, самата невронна мрежа не се променя, променят се само входните данни. Освен това такива „противникови примери” често остават ефективни и срещу невронни мрежи различни от тази, за

която са създадени, ако те също са обучени на същия набор от данни. Трябва да се отбележи, че ако добавеният шум е случайно генериран, негативният ефект върху точността на мрежата е значително по-малък. В следващи публикации е показано, че за ефективна атака на набора от данни MNIST, много прости модели, като например логистичната регресия са достатъчни за генериране на примери, които са ефективни срещу по-сложни архитектури. Докато началните атаки модифицират дигитално изображение съхранено в паметта и подадено като вход директно на невронната мрежа, [30] показва, че такъв тип атаки могат да са успешни дори, когато няма достъп до самият модел, а данните се събират през сензор – например камера на смартфон. Докато в ранните разработки на тази тематика се използва, че архитектурата на невронната мрежа и теглата ѝ са известни на атакуващия, по-късни разработки показват, че е възможно да се изработят атакуващи примери и без такава информация. Макар, че има разработки и в обратната насока – за засичане и предотвратяване на такива атаки, те все още не успяват да се справят с най-добрите демонстрирани алгоритми за създаване на атакуващи примери. [26] прави обзор на различни публикации на тема атака на невронни мрежи използвани за компютърно зрение с акцент върху приложението им срещу модели за засичане на характерните точки на лицето. Същевременно съществуват техники, които пречат на коректното засичане на лица чрез използване на стикери и аксесоари в реалния физически свят. В различни публикации е показано е, че при контролирани условия е възможно да се заблуди детектор за лица чрез използване на слънчеви очила със залепени стикери със специални шарки (фигура 10,а) или чрез специално добавени петна върху медицинска маска (фигура 10,б) [26]. В случай, че не може да бъде намерено лицето, няма как да се извърши и засичане на характерните точки на лицето. Съществуват и публикации фокусиращи се конкретно върху заблуждаване на алгоритми за намиране на характерните точки на лицето.



***Фигура 10**: Начини за целево противодействие на методите за засичане на характерни точки на лицето: а – слънчеви очила с добавени специални шарки; б - медицинска маска със специално добавени черни точки*
Източник: [26]

[26, 34] правят детайлен обзор на алгоритмите за засичане на характерните точки на лицето, от който вадят следните заключения:

1. Въпреки че в последните години има значително развитие при алгоритмите за засичане на характерни точки на лицето, много малко от разработките са фокусирани върху реалното приложение на тези алгоритми, поради което често, дори при изпълнение на видеокарта на настолен компютър, те не могат да работят в реално

време (за реално време се приема около 30 кадъра в секунда или 33 милисекунди на изображение).

2. Много приложения се изисква да работят ефективно върху мобилни устройства, но от разгледаните публикации само в една [21] оригиналните автори директно целят използване на модела им в мобилни приложения.
3. Въпреки, че най-съвременните разработки се фокусират върху работа с изображения събирани в неконтролирани условия, с голямо извъртане на главата, различни изражения и вариращa осветеност, каквито са набори от данни като 300W и AFLW, все още е лабо развитието на алгоритмите към още по-предизвикателните условия когато части от лицето са закрити. В последно време този проблем става все по-актуален с налагането на задължително носене на защитни маски.

3. Стандартни метрики и набори от данни

3.1 Метрики използвани за оценка и сравнение точността разработките в областта

Най-разпространената метрика за оценка на точността на алгоритмите за засичане на характерните точки на лицето е нормализираната осреднена грешка (normalized mean error, NME), представляваща евклидовото разстояние между предвидените и истинските позиции на точките, нормализирано спрямо даден коефициент за всяко лице, най-често разстоянието между очите. Тази метриката се дефинира така:

$$NME = \frac{1}{N}\sum_{i=1}^{N} NME_i;\ \ NME_i = \frac{1}{L}\sum_{j=1}^{L} \frac{\left\| x_{i,j}^{*} - x_{i,j} \right\|_2}{d} \qquad (16)$$

където N е броят на изображенията в набора от данни, L – броят на засичаните характерни точки x* - анотираните координати на характерните точки, x – предвидените такива, а d е нормализирашият коефициент. Целта на нормализиращия коефициент е всяко лице да дава равен принос в общата грешка без значение от резолюцията на изображението и частта от него заета от лицето. В различните набори от данни е прието да се използват различни коефициенти за нормализация, най-често използваният е разстоянието между очите, като обикновено се взима разстоянието между центровете на зениците, но някои публикации ползват разстоянието между ъгълчетата на очите. В по-скорошните набори от данни, които включват по-разнообразни ориентации на главата, се забелязва, че тази нормализация не работи добре при нефронтални лица, където това разстояние е значително по-малко отколкото при фронталните такива и дори клони към нула, когато лицето е в профил. Затова се въвеждат други коефициенти за нормализация – диагонала на описания около лицето регион или средно геометричното на дължината и ширината му. Тук отново има две вариации – някои ползват региона анотиран от детектора на лица, но това е съпътствано от проблема, че различните набори от данни използват различни детектори за анотация. Други публикации заобикалят този проблем използвайки

минималния описан около анотираните характерни точки правоъгълник. Често тази метрика се дава в проценти, дори това да не е указано изрично.

Друга важна метрика е процента на неуспехите (Failure Rate). Като неуспехи се определят тези изображения, за които нормализираната средна грешка е над даден праг. Той се дефинира като :

$$FR = \frac{1}{N}\sum_{i=1}^{N}\begin{cases}0, & NME_i < t \\ 1, & NME_i \geq t\end{cases} \qquad (17)$$

където t е избрания праг, а N броя снимки използвани за оценката. Различните публикации използват различни стойности за прага.

Площ под кривата на разпределението на грешката (cumulative error distribution – area under the curve, CED-AUC) – тъй като различните автори използват различни стойности на прага за процента на неуспехите, директно сравнение по този показател не може да се направи, затова е прието да се сравнева и съотношението на площа под кривата на разпределението на грешката. Това дава добра представа за устойчивостта на съответните модели към вариациите в набора от данни.

3.2 Публично достъпни набори от данни използвани като стандарт за за оценка и сравнение точността разработките в областта

В тази секция правим кратък преглед на различните публично достъпни набори от данни използвани за обучение и оценка на алгоритмите за засичане на характерните точки на лицето. Всеки от тях има собствен метод за разделяне на обучително и тестово множества, определя метрики за сравнение на алгоритмите и други детайли за начина на сравнение, които са описани в публикацията, където набора от данни е представен за първи път.

Таблица 2 показва обобщена информация за различните публично достъпни набори от данни – условия при които са заснемани изображенията, общ брой обучителни и тестови изображения, брой самоличност (където са обявени), брой характерни точки, с които са анотирани, и ориентация на главата (ъгъл на завъртане в хоризонталната равнина спрямо фронталната позиция). В някои набори от данни се срещат по повече от едно лица на едно по-голямо изображение, в който случай ги броим отделно. COFW-68 има само тестово множество (повече за това в секцията за него):

Набор от данни	Условия	Брой лица/изображения	Брой самоличност	Брой точки	Ориентация на главата
Multi-PIE	контролирани	~750000	337	68	[-45°, 45°]
XM2VTS	контролирани	2360	295	68	0°
FRGC-V2	контролирани	4950	466	5	0°
AR	контролирани	~4000	126	22	0°
LFPW	неконтролирани	1035	-	35	[-45°, 45°]

HELEN	неконтролирани	2330	-	194	[-45°, 45°]
AFW	неконтролирани	468	-	6	[-45°, 45°]
AFLW	неконтролирани	25933	-	21	[-45°, 45°]
AFLW-68	неконтролирани	25933	-	68	[-45°, 45°]
COFW	неконтролирани	1852	-	29	-
COFW-68	неконтролирани	507	-	68	-
IBUG	неконтролирани	135	-	68	-
WFLW	неконтролирани	10000	-	98	[-90°, 90°]

***Таблица 2**: Основни характеристики на различните набори от данни*
Източници: [4, 26, 34]

Multi-PIE: Наборът от данни Carnegie Mellon University - Multi Pose Illumination, and Expression (Multi-PIE) [36] съдържа около 750000 изображения на 337 индивида заснети в лабораторни условия при 4 различни сесии. За всеки индивид са налични изображения за 15 различни ориентации на главата, 19 нива на осветеност и 6 изражения (неутрално, писък, усмивка, присвиване на очите, изненада, отвращение). Анотирани са 68 характерни точки на лицето за всяко изображение (фигура 12, a) с завъртане на главата в диапазона [−45°, 45°].

XM2VTS: Наборът от данни Extended Multi Modal Verification for Teleservices and Security applications (XM2VTS) [36] съдържа 2360 фронтални изображения на 295 различни индивида, заснети при 4 различни сесии. За всеки индивид има по 2 снимки от всяка сесия. Всички индивиди са заснети при еднаква осветеност и повечето са с неутрално изражение. За всяко изображение са анотирани по 68 характерни точки (фигура 12, b), но точността на анотация в някои случаи е ниска, а позициите на точките са различни от тези в Multi-PIE.

FRGC-V2: Наборът от данни Face Recognition Grand Challenge Version 2.0 (FRGCV2) [36] съдържа 4950 изображения на 466 различни индивида, като за всеки индивид има изображения заснети в добре контролирани условия (равномерна осветеност, висока резолюция), както и изображения при лоши условия (неравномерна осветеност и ниско качество). Анотираните характерни точки са само 5 (фигура 12, c).

AR: Наборът от данни AR [36] съдържа над 4000 изображения на 126 индивида (70 мъже и 56 жени). Изображенията са заснети в две сесии за всеки индивид и се състоят от фронтални изображения с вариращи изражение, осветеност и частично закриване на лицето (слънчеви очила и шал). Анотирани са с 22 характерни точки (фигура 12, d).

LFPW: Наборът от данни Labeled Face Parts in the Wild (LFPW) [36] съдържа 1287 изображения свалени от интернет сайтове като google.com, flickr.com, yahoo.com и други. Този набор от данни предоставя само URL адресите, а не самите изображения. Авторите на [34] отбелязват, че са успели да свалят само 811 от 1100 обучителни изображения и 224 от 300 тестови изображения, поради неработещи адреси. Изображенията имат големи вариации в позицията на главата, изражението, осветеността и закриванията на части от лицето. Анотирани са 35 характерни точки на лицето (фигура 12, e) като според [34] точността при някои изображения е ниска.

HELEN: Наборът от данни HELEN [36] съдържа 2330 изображения свалени от уеб услугата flickr.com, които са на различни индивиди и имат големи вариации в позата, осветеността и израженията. Всяко изображение е с резолюция от приблизително 500 x 500 пиксела. Анотирани са много детайлно с 194 характерни точки (фигура 12, f), но точността на анотация е ниска.

AFW: Наборът от данни Annotated Faces in-the-wild (AFW) [36] съдържа 250 изображения с общо 468 лица, тъй като на много от включените изображения е анотирано повече от едно лице. Вариациите в изображенията с сходни с тези в гореизброените набори, като анотираните характерни точки са само 6 (фигура 12, g).

IBUG: Наборът от данни IBUG [36] е публикуван като част от първата версия на състезанието 300W. Състои се от 135 изображения свалени от интернет с големи вариации в израженията, осветеността и ориентацията на главата. Анотираните характерни точки спазват схемата използвана в Multi-PIE (фигура 12, a).

300W: Наборът от данни 300W [36, 37] съдържа колекция от различни предходни набори от данни като HELEN, LFPW, AFW и IBUG, които са преанотирани с 68 характерни точки на лицето (фигура 11, a). Изображенията са разделени в две групи – заснети на открито и заснети на закрито, с по точно 300 във всяка група. В [37] е указан начина по който да се разделят изображенията на обучаващи и тестови. Тестовия набор е разделен на обикновен набор, предизвикателен набор и пълен набор (предходните два комбинирани). Често публикациите указват нормализираната средна грешка за всеки от тези поднабори по отделно. Коефициентът за нормализация (d във формула (16)) е указан като разстоянието между центровете на зениците, с цел лицата с различни размери да имат еднакъв принос в крайната грешка. Изображенията в 300W за заснемани при различни условия (осветеност, цветова гама, емоции) и съдържат лица заснети под ъгъл.

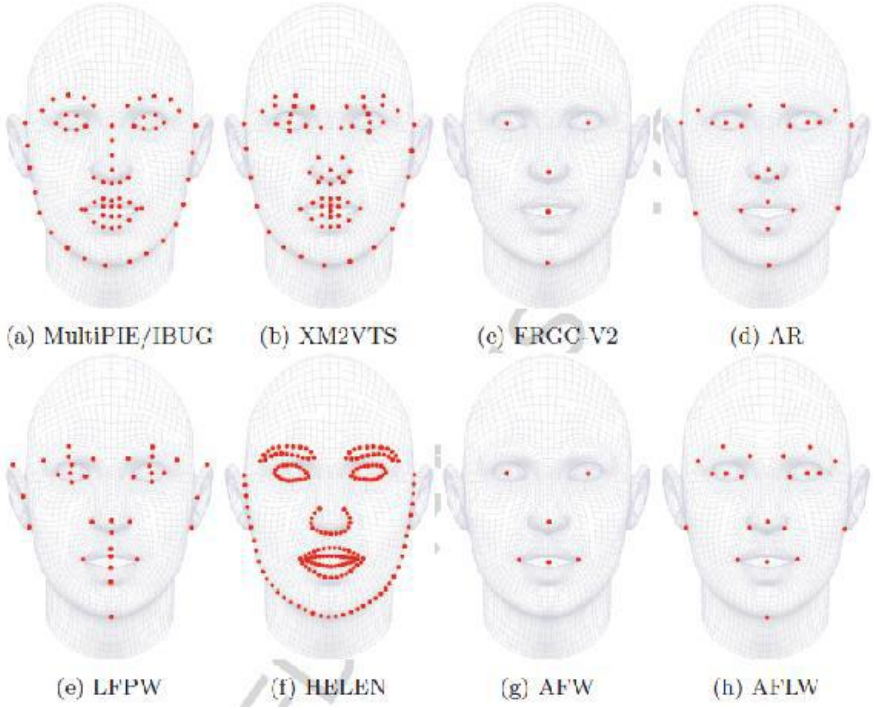
AFLW: Наборът от данни Annotated Facial Landmarks in the Wild (AFLW) [36] съдържа 25993 изображения събрани от flickr.com, имащи големи вариации в ориентацията на главата, изражението, възрастта, пола, етническата принадлежност и други условията на заснемане като цяло. Ъглите на заснемане са по-големи от тези в предходните набори от данни, като авторите му предлагат да се раздели на два поднабора - фронтален и пълен. Анотирани са 21 характерни точки на лицето (фигура 11, б; фигура 12, h), като има и преанотирана версия с 68 характерни точки наречена **AFLW-68** [35], но използвана по-рядко в практиката. В [29] е представен наборът от данни **MERL-RAV**, който съдържа изображенията от AFLW преанотирани с 68 характерни точки, с добавени етикети за видимостта на всяка точка – видима, самозакрита (от ориентацията на главата) или закрита от друг обект.

COFW: Наборът от данни Caltech Occluded Faces in the Wild (COFW) [5] е по-предизвикателен (фигура 11, в), като се фокусира върху анотиране на изображения, които са частично закрити от други обекти по естествен начин – например микрофон, спусната коса, ръка пред устата и т.н. Този набор указва като метрики за сравнение не само нормализираната средна грешка (NME), но и процента на провалите (означен с FR - уравнение 17) – изображенията, при които грешката е над даден праг. За този набор също съществува преанотирана версия с 68

характерни точки - **COFW-68** [18], които цели да се използва за оценяване с по-предизвикателни изображения на алгоритми, обучени върху някой от предходните набори използващи този формат на анотация.

WFLW: Наборът от данни Wider Facial Landmarks in-the-wild (WFLW) [47] е един от най-новите и най-трудни, тъй като съдържа изображения с големи вариации на показваните емоции, ориентацията на главата, използван грим, осветеността, както и размазани изображения (фигура 11, г), а също така и е анотиран много детайлно – с 98 характерни точки. Указаните метрики за използване са три – нормализирана средна грешка (NME), процент на провалите (FR) с определен праг, както и площ под кривата на разпределението на (CED-AUC).

***Фигура 11**: Примерни изображения от наборите от данни:*
а – 300w; б – AFLW; в – COFW; г - WFLW
Източник: [26]



***Фигура 12**: Конфигурации на характерните точки на лицето в някои от разпространените набори от данни.*
Източник: [34]

[26, 34] правят обзор на част от разгледаните по-горе набори от данни, както оценка на някои по-популярни скорошни разработки върху тях. [4] предлага модел, предвиждащ триизмерните координати на характерните точки (добавят дълбочина към пикселните координати), като за целта изграждат набор от данни комбиниращ изображенията на част от разгледаните набори и разширяващ тяхната анотаци да включва дълбочина. Предоставят детайлни стравнения в точността на разработения модел върху използваните набори от данни, както в модифицираните от тях версии с 3D анотации, така и в оригиналните версии.

4. Предишни разработки

В тази глава ще направим кратък обзор на по-значимите разработки в областта, използваните от тях алгоритми и архитектури на невронни мрежи, както и по-важните им характеристики, без да правим оценка на точността им и да даваме имплементационни детайли, каквито могат да бъдат видяни в оригиналните им публикации.

Ранните разработки в областта са главно основани на напасване на деформируем модел на лицето, като най-значимите такива алгоритми са Active Appearance Model (AAM) и Active Shape Model (ASM), които разгледахме в секции 2.31 и 2.3.2. Те изчисляват позициите на характерните точки на лицето въз основа на напраснатия модел на формата на лицето. В повечето случаи тези алгоритми използват статистически методи в основата си. Те имат достатъчно добра точност в контролирани условия (фронтално разположено лице и добра осветеност), но приложимостта им в реалния свят е доста ограничена, тъй като при изображения, заснети в неконтролирани условия, точността им много се влошава. Следващите ги разработки ползват методи основани на техниките random forests и gradient boosting, като например алгоритъмът ERT [25], който ще разгледаме първи в тази глава. Тези разработки имат по-добра точност, но също се провалят при определени условия на заснемане на изображенията.

Към момента алгоритмите, основани на невронни мрежи, дават най-ниска грешка при засичане на характерните точки на лицето в изображения с голям ъгъл на заснемане и частично закриване на големи части от лицето. Тези алгоритми включват методи с директна регресия, които предвиждат директно x и y координатите на всяка характерна точки, методи с регресия на „топлинни карти” (heatmaps), които изграждат 2D топлинна карта за всяка характерна точка на лицето. Стойностите на тези карти могат да се разглеждат като вероятности съответната характерна точка да се намира на дадени координати в изображението. Някои алгоритми също така използват каскади от регресори, където предвижданото за координатите се подобрява на няколко стъпки.

Dlib [27] е библиотека за машинно самообучение с отворен код. Сред включените в нея алгоритми е и този за засичане на характерни точки на лицето наречен ERT [25] който е каскаден регресор използващ gradient boosting. При работата на ERT характерните точки се локализират чрез итеративно напасване на шаблон на лицето, който е образуван чрез осредняване на обучителните данни и се инициализира спрямо описания около лицето правоъгълник резултат от детектор за лица на Viola-Jones. Основното предимство на ERT е високата скорост при засичане на

характерните точки на лицето (според авторите на алгоритъма около 1 милисекунда на лице). Имплементацията на алгоритъма включена в библиотеката е обучена върху набора от данни 300W. Този алгоритъм се използва широко при съвременните проучвания в областта поради високата си скорост на работа и наличието на имплементации с отворен код. Все пак по-новите публикации показват, че използването на невронни мрежи е за предпочитане при лица с големи вариации в ориентацията на главата, поради много по-висока точност в тези случаи [17].

Multi-task Cascaded Convolutional Networks (MTCNN) [52] е подход, при който невронните мрежи се обучават съвместно да засичат самите лица, както и характерните им точки (в оригиналната публикация конкретно 5 точки – очите, върха на носа и краищата на устата), което подобрява точността и на двете задачи. Самата невронна мрежа се изгражда като каскада от три мрежи - Proposal Network (P-Net), Refine Network (R-Net), Output Network (O-Net). Всяка от тях предвижда описан правоъгълник за лицето, вероятността, че в този правоъгълник има лице и петте характерни точки. P-Net е бърза изцяло конволюционна мрежа, която обработва оригиналното изображение в множество резолюции (така наречената пирамида от изображения). Тази мрежа произвежда множество груби предвиждания за описани около лицата правоъгълници, които се филтрират с алгоритъм Non-Maximum Suppression (NMS). След това R-Net мрежата подобрява точността на предвидените правоъгълници, без да обработва отново цялото изображение, за да намали времето за изчисления. След нея отново се прилага филтриране с NMS. Накрая O-Net мрежата прави финалното уточняване на описаните правоъгълници. Тази мрежа е най-бавната от трите, но обработва само малък брой правоъгълници. Важна характеристика на тази разработка е, че прави подбор на трудните примери по време на обучение, при което мрежата се обучава на предизвикателните примери, докато тези, на които предвиждането ѝ вече е достатъчно точно, се пропускат. В оригиналната публикация авторите използват около 70% от най-трудните примери от всяка група използвана за обучение.

Dense Face Alignment (DeFA) [33] е единственият алгоритъм разгледан в тази глава, където невронната мрежа се използва за предвиждане на характерните точки на лицето чрез деформируем 3D модел на лицето. Интересното в този алгоритъм е, че позволява да се изгради плътна 3D мрежа използвайки само 2D изображения, която може да отрази голям набор от ориентации на главата и изражения на лицето. Интересно е също и, че DeFA може да бъде директно обучен върху набори от данни с различен брой анотирани характерни точки, тъй като те се отчитат като ограничения за изграждания 3D модел.

Style Aggregated Network (SAN) [14]. Авторите на тази разработка отчитат вариациите в стила на изображението от наборите от данни 300W и AFLW, които могат да бъдат тъмни или светли, цветни или черно-бели и т.н. Предходните разработки не взимат тази информация под внимание. Авторите забелязват, че в зависимост от стила на изображението, предходните алгоритми дават различни отклонения в засечените позиции на характерните точки на лицето, с по-големи грешки при изображения с по-екстремна осветеност. Решението, което те предлагат е първо да се обучи невронна мрежа от тип Generative Adversarial Network Cycle (GAN), която да трансформира изображенията към неутрален стил, а след това да се обучи отделна невронна мрежа, която да предвижда характерните точки на лицето използвайки за вход и двете

изображения – „неутралното” и оригиналното, тъй като в „неутралното” може да липсват финни детайли, които в някои случаи оказват влияние не точността.

Look at Boundary (LAB) [47]. Главният принос на авторите на тази разработка е, че предлагат топлинни карти (heatmaps) на контурите на частите на лицето (очи, нос, скули, брада и т.н.), които се създават като междинно представяне между оригиналното изображение и предвидените характерни точки на лицето. Този подход подобрява качеството на предвижданията и същевременно позволява първия модул, който търси контурите, да бъде обучаван на набори от данни с различни схеми на анотация, докато вторият модул се обучава по отделно за всеки набор. Авторите са показали, че обучаването на първия модул върху набора от данни 300W подобрява финалните предвиждания на характерните точки върху наборите от данни AFLW и COFW. В същата публикация авторите предлагат и нов, по-предизвикателен набор от данни наречен от тях (фигура 11, г).

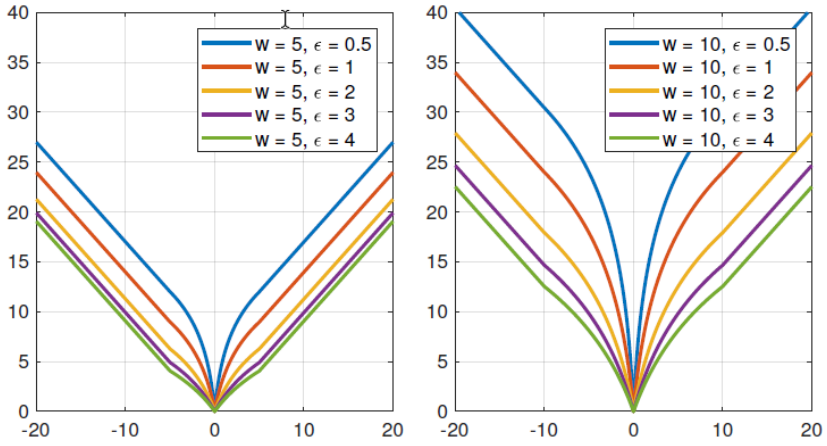
Wing Loss [17]. В тази разработка авторите забелязват, че влиянието на функциите на грешката върху обучението на невронни мрежи за засичане на характерните точки на лицето е слабо проучен въпрос. Повечето разработки $L2 = \frac{x^2}{2}$ като функция на грешката за директната регресия на характерните точки, която е известно, че е силно чувствителна към изключения (outliers), което някои по-ранни публикации заобикалят, използвайки гладка L1 функция на грешката (уравнение 18). Авторите на тази публикация правят сравнение на L2 функцията с различни функции на грешката, като L1 и гладка L1.

$$smooth\ L1(x) = \begin{cases} \frac{x^2}{2}, & |x| < 1 \\ |x| - \frac{1}{2}, & |x| \geq 1 \end{cases} \qquad (18)$$

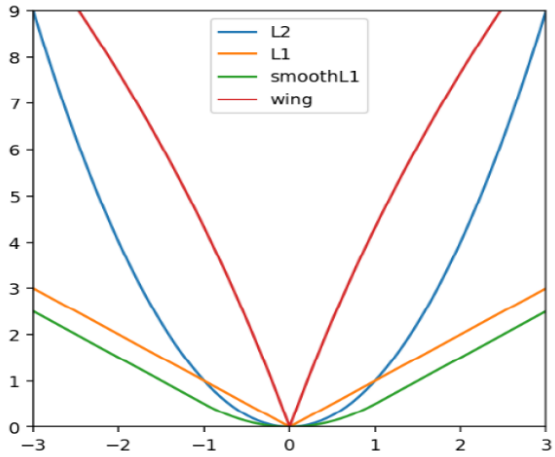
и отчитат, че те дават по-добри резултати. Главният принос на тази разработка е, че предлага нова функция на грешката, наречена Wing loss (уравнение 19), която комбинира L1 за големи отклонения в позициите на характерните точки и натурален логаритъм за малки отклонения. Името е вдъхновено от графиките на функцията при различни стойности на хиперпараметрите, приличащи на криле (фигура 13).

$$wing(x) = \begin{cases} w\ ln(1 + \frac{|x|}{\epsilon}), & |x| < w \\ |x| - C, & |x| \geq w \end{cases} \qquad (19)$$

където $C = w - w\ ln(1 + \frac{w}{\epsilon})$, w и ϵ са хиперпараметри (w=15,ϵ=3 в оригиналната публикация). В допълнение за да се обучава мрежата повече върху трудни примери, редките обучителни примери (според ориентацията на главата) се дуплицират с изменения.



***Фигура 13**: Графики на функцията на грешката Wing loss за различни стойности на хиперпараметрите. а – w=5; б – w=10*
Източник: [17]



***Фигура 14**: Сравнение на функциите на грешката ℒ2, ℒ1, smoothL1, Wing loss (при и=15, ϵ=3).*
Източник: [26]

Practical Facial Landmark Detector (PFLD) [21]. Тази публикация отчита по-добър резултат според метриката за нормализирана средна грешка (NME) върху наборите от данни 300W и AFLW, като същевременно предлага прост за имплементиране алгоритъм, който авторите целят да се прилага директно в мобилни приложения. Авторите твърдят, че към момента на публикацията, това е единственият съвременен алгоритъм за засичане на характерните точки на лицето, основан на невронни мрежи, който може да работи в реално време на мобилни устройства. В основата на разработката си те използват архитектура MobileNetV2, представена оригинално в [38] с цел класификация на изображения, за извличане на характеристики от изображенията. Към нея те добавят две разклонения – едно за засичане на характерните точки на лицето, използваща напълно конволюционни слоеве с няколко резолюции и една с опростен 3D модел на главата за изчисляване на ъглите на завъртането ѝ (yaw, pitch и roll – фигура 9). Второто разклонение се използва само при обучение на мрежата и се състои от няколко конволюционни слоя.

AWing [46]. Този алгоритъм използва регресия върху топлинни карти (heatmaps), като за всяка характерна точка на лицето създава карта с размер 64×64, върху която точната позиция на точката се определя с използване на филтър с гаусова дистрибуция с размер 7×7. Тази разработка комбинира идеи от предшестващите я публикации [4], [17](разгледаната по-горе Wing Loss), [47] (разгледаната по-горе Look at Boundary) и [32]. Авторите са забелязали, че при използване на L2 за функция на грешката, създадените топлинни карти нямат достатъчно силно изразени характеристики при по-предизвикателни изображениям тъй като тя е слабо чувствителна към малки грешки, а в оригиналната си форма Wing loss не е подходяща за регресия върху топлинни карти, тъй като производната ѝ е прекъснат в нулата. В допълнение към това, за всяка топлинна карта съществува проблема за дисбаланс в класовете, тъй като само няколко пиксела имат положително значение (означавайки, че е вероятно съответната характерна точка да се намира там), докато по-голямата част от пикселите се отбелязват с отрицателно значение, което също не е отчетено в оригиналната имплементация на Wing loss функцията. За да компенсират всичко това авторите предлагат функцията Adaptive Wing loss, която е диференцируема около нулата и подчертава малките грешки в пикселите с положително значение, ни не и в тези с отрицателно.

В допълнение към разгледаното в тази глава, [6] прави подробен преглед, оценка и сравнение на някои по-известни разработки спрямо типа на използвания алгоритъм. [24] прави преглед и сравнение на скорошните разработки фокусирани върху работа с изображения заснети при неконтролирани условия, а [7] се концентрира върху сравнение на разработки, основани само на дълбоки невронни мрежи. [26] използва набора от данни 300W, за да сравни точностите на множесто съвременни разработки, като в допълнение прави и сравнение на тяхната скорост, както при изпълнение върху процесора, така и върху видеокарта.

5. Разработено решение за засичане на характерни точки на лицето

5.1 Избрана постановка на задачата

Разгледаните в секция 2.1 приложения на задачата за засичане на характерни точки на лицето и нейната недвусмислена формална постановка, ясно определят изискванията към работата на един алгоритъм за засичане на характерни точки на лицето - по подадени изображения с вече намерени в тях лица и описаните правоъгълници около тези лица, за всяко едно лице да генерира списък с пикселни координати за определен брой характерни точки на лицето.

Броят точки, тяхната номерация и очакваните им позиции са различни в различните набори от данни. Различни са също така и използваните от различните разработки детектори на лица и съответно техните пристрастия (bias), което води до различно разположение на характерните точки спрямо границите на описания правоъгълник. От разгледаните в секция 3.2 набори от данни и в глава 4 – значими предишни разработки, става ясно, че всички скорошни разработки се фокусират върху работа с изображения събирани при неконтролирани услови. Най-

често използвания за сравнение на точността между различни алгоритми набор от данни в тази категория е 300W, а най-често използваната метрика е нормализираната средна грешка (normalized mean error, NME – уравнение 16). 300W включва анотации с 68 характерни точки на лицето във формата въведен от по-ранни набор от данни Multi-PIE. 300W не оказва конкретен детектор на лица, който да бъде използван и анотациите не включват описаните около лицата правоъгълници. Те обаче могат да бъдат създадени за всяко лице чрез разширение на минималния описан правоъгълник, съдържащ характерните му точки, така, че добре да обхваща цялото лице. За определеност приемаме разширение с 30% съответно от дължината или ширината във всяка от четирите посоки.

Съгласно описаните в секция 1.2 цели на дипломната работа, изискване към ефикасността на разработения алгоритъм е да може да работи в реално време на вградени системи. Една конкретна платформа за вградени системи е Nvidia Tegra TX2. Тя е проектирана конкретно с цел цел вграждане на решения основани на изкуствен интелект в мобилни и вградени устройства и е широко използвана в автомобилната индустрия. Така практическата постановка на нашата задача става да се създаде модел за засичане на характерните точки на лицето, постигащ оптимална точност върху набора от данни 300W, спрямо метриката нормализирана средна грешка, постигайки скорост на работа от поне 25 кадъра в секунда върху платформата Nvidia Tegra TX2.

От секция 3.2 става ясно, че макар да е най-широко използваният за сравнения набор от данни, 300W е сравнително малък по размер и с доста по-ограничени вариации на условията на заснемане на изображенията в сравнение с по-късните набори от данни. Един от най-предизвикателните и големи по размер набори е WFLW (също разгледан в секция 3.2). Неговата анотация включва 98 характерни точки на лицето, но техният формат е такъв, че те надграждат 68-те точки използвани в 300W. Поради това можем да използваме и наборът от данни WFLW за обучение на модела, като изключим част от анотираните точки.

5.2 Метод за решаване на задачата

5.2.1 Използвани алгоритми и архитектури на дълбоки невронни мрежи

От разгледаните в глава 4 предходни разработки става ясно, че от факторите, причиняващи значителни вариации във външния вид на лицето върху изображенията и съответно успеха на засичане на характерните точки, най-голямо влияние оказва ориентацията на главата, тъй като тя причинява както големи разместванията в относителната позиция на точките, така и самозакриване на части от лицето. Предходните разработки прилагат два различни подхода за справяне с този фактор – чрез предварително изчисление на грубо приближение на ориентацията на главата и изграждане на гора от отделни модели, всеки съответстващ на дадена ориентация или чрез отчитане на ориентацията на главата във функцията на грешката и придаване на повече тежест при обучение на примери с по-предизвикателна ориентация. Най-добра постигната точност заявяват разработките използващи първият подход, но те са и най-тежки за изпълнение, като скоростта им е далече от необходимата за изпълнение в реално време дори и върху видеокарта на настолен компютър. Поради това ние избираме втория подход.

Като част от предварителната обработка на обучителните данни, ние ще използваме анотираните характерни точки на лицето и грубо приближение на техните координати при фронтално разположено лице, прилагаме Perspective-n-Point solver (PnP) от библиотеката за компютърно зрение OpenCV, за да разширим анотацията на данните с грубо приближение на ойлеровите ъгли, показващи ориентацията на главата. По време на обучение прилагаме същия метод за изчисляване на ъглите съответстващи на изчислените характерни точки. Разликата между двете групи ъгли използваме като коефициент за да умножим стандартната L2 метрика за грешка, по следния начин:

$$Err = \frac{1}{M}\sum_{i=1}^{M}\left(Max\left(0.1, \sum_{j=1}^{3}(1-\cos\theta_{i,j})\right)\sum_{k=1}^{L}\left\|x_{i,k}^{*} - x_{i,k}\right\|_2^2\right) \qquad (20)$$

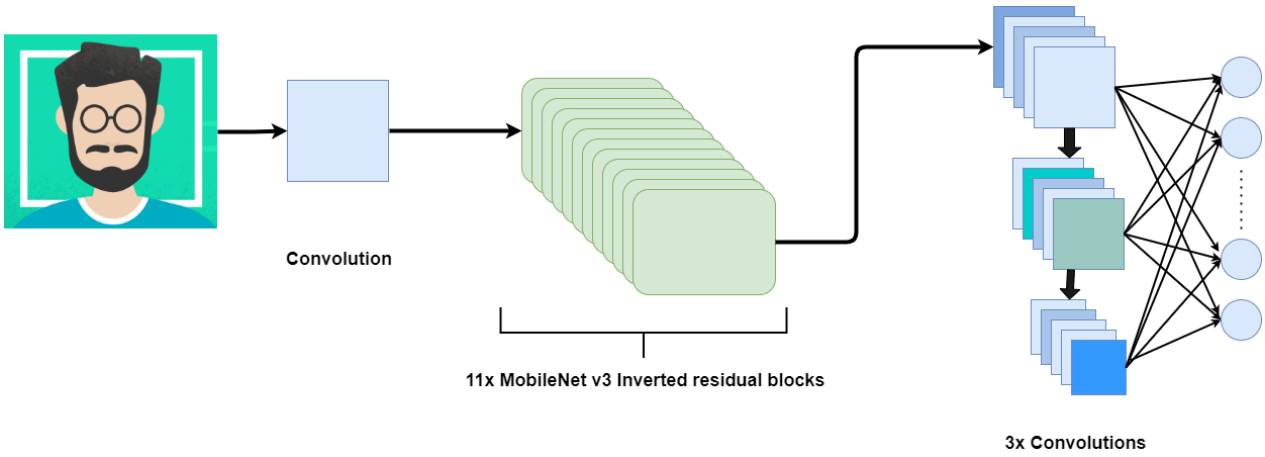
Където M е броят изображения в съответната група за обучение, L е броят на засичаните характерни точки – в нашият случай 68, $\theta_{i,j}$ е разликата между съответните ъгли изчислени на база анотираните точки и засечените точки, $x_{i,k}^{*}$ е анотираната позиция на характерната точка k на изображение i, а $x_{i,k}$ е изчислената ѝ позиция. За да може лицата, чиято истинска ориентация е близка до изчислената от засечените характерни точки, да не бъдат пренебрегвани от функцията на грешката. Това е много важно, тъй като голяма част от лицата в използваните набори от данни са с фронтална или близка до фронталната ориентация и без това ограничение техният принос бързо би намалял до несъществен.

Авторите на [21] показват, че използвайки за основа архитектура на невронна мрежа предназначена за класифициране на изображения, но проектирана за работа на мобилни устройства, постигат изключително добър баланс между скорост на работа и точност, като към момента на публикацията са с най-висока точност спрямо метриката нормализирана средна грешка (NME) сред моделите, постигащи работа в реално време на мобилни устройства. Те използват за основа архитектурата MobileNetV2 [38], ние ще използваме по-новата и постигаща по-добра класификационна точност и скорост на работа MobileNetV3Small [22].

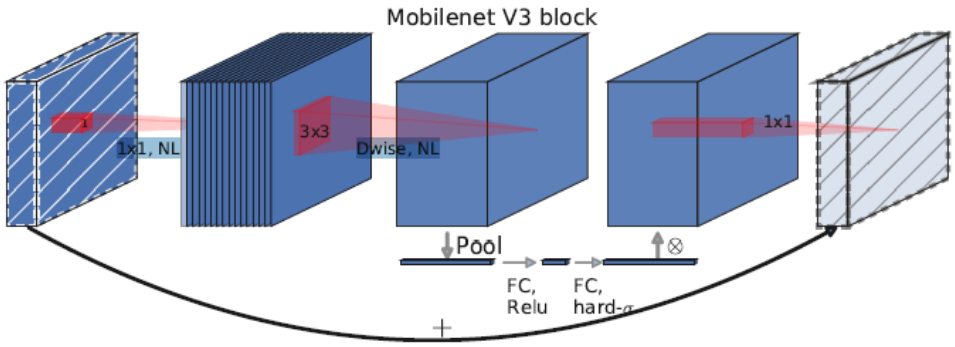
Използвайки тази архитектура за извличане на характеристиките на изображенията, ние я разширяваме, като взимаме в предвид, че в частният случай (спрямо общата задача за извличане на характеристики от изображения) на работа със снимки на човешки лица в изображенията е налична силно изразена глобална геометрична структура – симетрия и строга пространствена свързаност между очите, носа, устата и контурите на лицето. Затова в края на архитектурата вместо карти на характеристиките с един мащаб (single scale feature maps) ние добавяме няколко различни мащаба, следвани от конволюции с отместване и накрая напълно свързан слой, който служи като връзка между различните мащаби.

Така в крайната архитектура на модела ние прилагаме 11 блока от тип Inverted residual заимствани от MobileNetV3Small, след които добавяме три двуизмерни конволюции с размери на ядрата съответно 1x1, 3x3 и 7x7, всяка следвана от GlobalAveragePool слой и Flatten слой за преобразуване на изходите им в едноизмерни масиви и обединение в едно чрез Concat слой. Последният слой на мрежата е напълно свързан – Dense, получаващ изходите и от трите конволюционни групи и произвеждащ крайните предвиждания за позициите на характерните

точки. Цялостната архитектура е представена условно на фигура 15, а архитектурата на блоковете от тип Inverted residual – на фигура 16.



***Фигура 15**: Схема на използваната архитектура.*



***Фигура 16**: Схема на Inverted residual блок от MobileNetV3 архитектура.*
Източник: [22]

5.2.2 Интерфейс на крайния модел

Крайният модел е имплементиран на езика Python чрез библиотеката за изкуствен интелект TensorFlow версия 2.4.0. За използване на модела извън ограниченията на библиотеката TensorFlow, след приключване на тренирането го конвертираме към стандартния формат за представяне на невронни мрежи с отворен код – Open Neural Network Exchange (ONNX), наложил се като стандарт при обмена на невронни мрежи. За изпълнение върху платформата за вградени системи Nvidia Tegra TX2 зареждаме конвертирания в ONNX формат модел в софтура за изпълнение на невронни мрежи Nvidia TensorRT.

Сериализираният във формат ONNX модел получава за вход серия от цветни изображения с размери 224 на 224 пиксела, представени като масив от 32 битови числа с плаваща запетая (float32) с размерност (n, 224, 224, 3), където n е броя на изображенията в групата, а 3 броят на цветните канали. Изходът от изпълнението му е серия от едноизмерни масиви – по един за всяко

подадено изображение, всеки с 136 елемента от 32 битови числа с плаваща запетая (float32), представляващи координатите на засечените точки. Четните позиции (започвайки индексирането от 0) представляват x координатите, а нечетните – y. Самите числа са в интервала от 0 до 1 и представят отношението на координатите съответно към ширината и височината на изображението. При сериализацията на предоставения с кода на дипломната работа модел използваният размер на групата е 1, с цел по-лесното му тестване върху индивидуални изображения. Използването на по-големи групи може да доведе до известно подобрение в скоростта на обработка.

5.3 Дизайн на експеримента

5.3.1 Обучаващо множество

За обучение ще използваме наборът от данни WFLW, тъй като е по-голям и по-разнообразен от 300W. Той е разделен от авторите му на обучително и тестово под множества, като първото съдържа 7500 лица, а второто – 2500 лица. Тъй като този набор от данни включва по-предизвикателни условия на заснемане от 300W, макар целта ни да е тестване на 300W за обучение ще използваме само съответното обучително подмножество. Към него ще добавим половината от изображенията в 300W, за да отчетем отчетем при обучението евентуалните особености в анотацията на набора от данни. Тъй като и двата набора от данни са анотирани ръчно, като и над двата са работили множество хора, отклоненията в анотациите се очаква да са случайни и за това се надяваме разликата в размерите на двата набора да не окаже особено влияние при обучението. Изображенията в 300W са разделени на заснети на открито и заснети на закрито, като размерите на двете групи са равни и вътре в всяка група изображенията са номерирани започвайки от 1, но тяхната подредба няма някакво конкретно значение. Затова за определеност ще подберем от всяка група нечетните изображения за обучение, а четните – за тестване.

За допълнително увеличаване на размера и разнообразието на обучителното множество, всяко едно изображение мултиплицираме, като изберем случайна точка около центъра на ограждащия лицето правоъгълник (координатите избираме, като към центъра на правоъгълника добавим случайно отклонение в интервала от -10% до +10% от ширината или височината съответно) и го ротираме около тази точка на случайно избран ъгъл в интервала [-20°, 20°]. По този начин допълнително симулираме различни ориентации на главата или ъгли на заснемане. Допълнително половината от ротираните изображения, избрани на случаен принцип, обръщаме огледално. Някои от разработките разгледани в глава 4 добавят и изкуствено закриване (добавяне на черни правоъгълници) на части от лицето, но тъй като в WFLW голямо количество снимки на хора носещи маски, слънчеви очила или други аксесоари, както и повечето снимки не са фронтални и съответно съществува естествено закриване на части от лицето, считаме това за излишно. За изображенията взети от WFLW прилагаме описанта мултипликация по 5 пъти, а за тези от 300W, поради по-малкият им брой – по 20 пъти.

След като приложим това мултиплициране, прилагаме и Perspective-n-Point solver (PnP) от библиотеката за компютърно зрение OpenCV, за да разширим анотацията на данните с грубо приближение на ойлеровите ъгли, показващи ориентацията на главата.

5.3.2 Процес на обучение и избрани хиперпараметри на модела

От обучаващото множество от данни отделяме на случаен принцип 25% за валидация. Използваме оптимизатора AdamW, с параметри learning rate – 10^{-4} и weight decay – $10^{-6.}$ За размер на групата за обучение на всяка стъпка (batch size) използваме 64, като това е съобразено с ресурсите на използваната за обучението система, разполагаща с видеокарта Nvidia GeForce GTX 1060 със 6GB памет. В рамките на всяка епоха за обучение използваме всички данни от разширения както е описано в горната секция набор от данни, като в допълнение по случаен начин намаляваме или увеличаваме яркостта на всяко изображение със стойност в интервала (-0.2, 0.2) от оригиналната му яркост.

Моделът обучаваме в продължение на 200 епохи, в края на които той продължава да показва подобрение в нормализираната средна точност върху валидационния набор от данни, но то е минимално. Това показва, че при наличие на повече ресурси и време за трениране може да бъде извлечена още малко допълнителна точност от същите архитектура и набор от данни.

5.3.3 Метрики за оценка

Съгласно избраната постановка на задачата ще разглеждаме метриката нормализирана средна грешка (NME – уравнение 16). Тъй като наборът от данни 300W оригинално е публикуван като част от състезание, възприетият начин за тестване върху него е използвайки целият набор, съответно това ще е водещата ни метрика. Като коефициент за нормализация ще използваме разстоянието между външните ъгълчета на очите. За пълнота ще разгледаме и нормализираната средна грешка по отделно за половината от снимките оставени за тест, както и по отделно за двете подмножества в набора от данни – заснетите на закрито и заснетите на открито, като вторите се считат за по-предизвикателни, както поради по-голямата вариация в осветеността, така и тъй като са събрани предимно от спортни събития и показват по-широк диапазон от емоции. Допълнително ще разгледаме и грешката върху тестовото множество на WFLW.

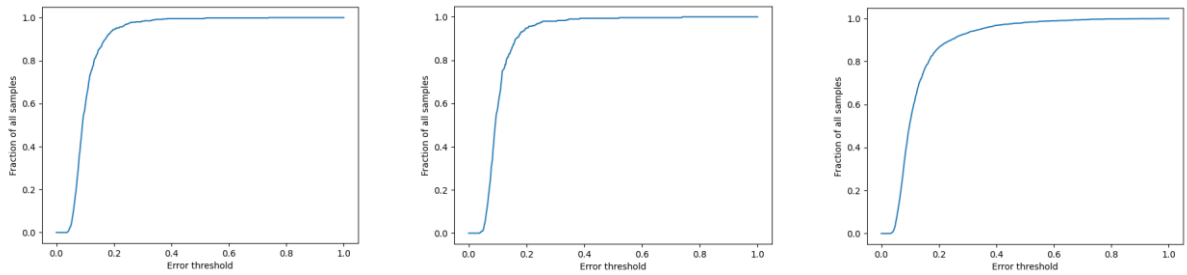
5.4 Резултати от експерименти

Разработеният модел постигна точност, спрямо метриката нормализирана средна грешка от 10.7% върху целият набор от данни 300W и 10.8% върху частта от набора 300W, която не е включена в обучаващото множество. Върху тестовото множество от данни на набора WFLW постигнатата точност е 13.1%, като от него са ползвани само 68 точки, съвпадащи като формат с тези от 300W, а не оригиналните 98. За сравнение в таблица 3 са дадени, постигнатите точности от разгледаните в глава 4 значими предходни разработки, според заявеното от авторите им. В допълнение показваме и кривите на грешката на разработения модел върху трите тестови множества (фигура 17). Постигнатите нива на метриката площ под кривата на грешката (cumulative error distribution – area under the curve) са съответно за целия набор 300W – 89.2%, за частта от набора 300W, която не е включена в обучаващото множество – отново 89.1%, за WFLW – 86.8%.

Скоростта на изпълнение върху платформата Nvidia Tegra TX2, при използване на софтуера TensorRT е 18.6 милисекунди за обработка на лице, при подаване на единични кадри към модела (batch size 1). Това е равно на обработка на 53 изображения в секунда и надхвърля изискванията за работа в реално време. При изпълнение върху настолен компютър с видеокарта Nvidia GeForce GTX 1060 (машината използвана за обучение на модела) постигнатата скорост при подаване на единични кадри е 2.6 милисекунди, равнозначно на 384 кадъра в секунда.

Име на разработения алгоритъм	NME (%)
ERT	6.40
DeFA	6.10
SAN	3.98
LAB	3.49
Wing Loss	4.04
PFLD	3.40
AWing	3.07

***Таблица 3**: Резултати на разгледаните предходни разработки върху набора от данни 300W.*
Източници: [26, 17]

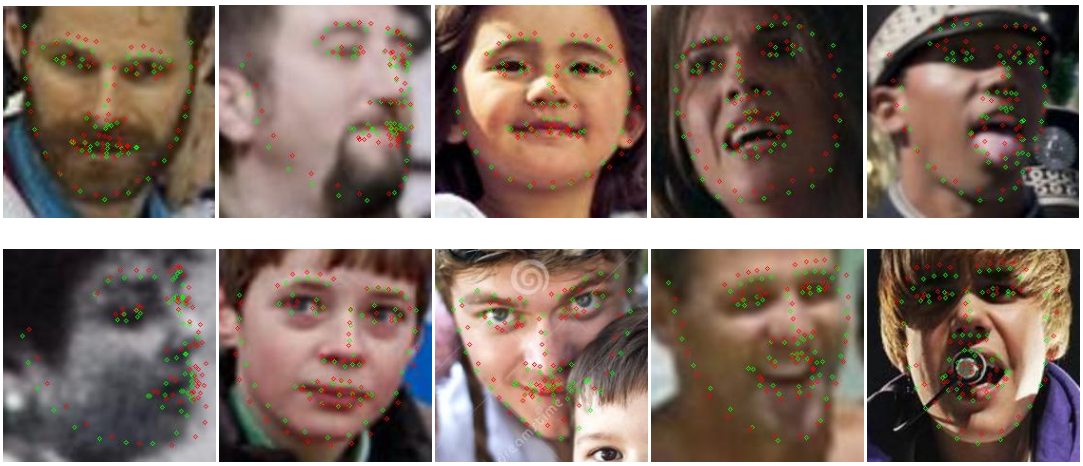


***Фигура 17**:Графики на кривата на грешката за разработената система.*
(от ляво на дясно – за 300W, за частта от набора 300W, която не е включена в обучаващото множество, за WFLW)

5.5 Примери от работата на разработената система

В тази секция показваме няколко случайно подбрани примера от работата на разработената система върху двата използвани набора от данни. Със зелено са означени анотираните позиции на съответните характерни точки на лицето, а с червено – засечените такива.

***Фигура 18**: Примери от работата на разработената система върху набора от данни 300W.*



***Фигура 19**: Примери от работата на разработената система върху набора от данни WFLW.*

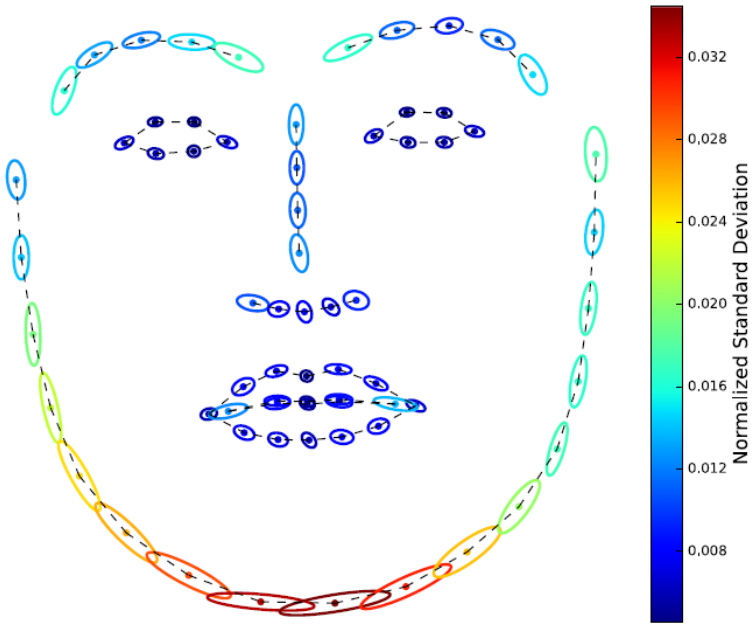
6. Заключение и бъдеща работа

Постигнатата от модела скорост от 53 кадъра в секунда при изпълнение върху платформата Nvidia Tegra TX2 отговаря и дори надхвърля на поставеното изискване за работа в реално време при прилагане във вградени системи. Постигнатата точност е далеч от водещите разработки в областта, но това е обяснимо с няколко фактора. На първо място е по-простата структура и по-малкото параметри на модела – необходими условия за постигането на работа в реално време вър системи с ограничени изчислителни ресурси, каквито представляват вградените системи. В допълнение разработеният модел е трениран на комбинация от два набора от данни. При тези от разгледаните предходни разработки, които посочват резултати върху повече от един набор от данни, се вижда, че най-добри резултати се постигат когато обучението и тестването е извършвано върху един и същ набор. Тогава обаче, моделите постигат и най-малка генерализация, като тези обучавани само върху 300W, често постигат незадоволителни резултати при тестване върху по-големи набори от данни. Освен това поради ограниченото време за

разработка модела не е обучаван до достигане на плато в метриката за точност. Тестваният модел е обучен ан 200 епохи, като това отнема над 24 часа. Предвид това, считаме резултатите постигнати от разработения модел за засичане на характерните точки на лицето за задоволителни.

Наблюдава се разлика между валидационната точност по време на обучение на модела, която достига до 12.8% и постигнатата точност върху тестовото множество – 10.8% върху 300W и 13.1% върху тестовото подмножество на WFLW. Това потвърждава, че модела е далече от overfitting върху тестовите данни и при по-дълго трениране може да постигне още по-добри показатели. Освен това наличието на по-разнообразни за обучение може допълнително да подобри точността на модела. Един подход за изкуствено генериране на повече данни, чрез използване на един единствен набор, използван в някои от разгледаните разработки, е налагане на анотираните точки върху 3D модел на главата и използването му за изкуственото генериране на варианти на всяко изображение с различни ориентации на главата. Това изкуствено разширение на обучителните данни е полезно и по друга причина. Немалка част от анотираните точки в използваните набори от данни се намират по контура на лицето и при тях точността на анотиране е по-ниска (фигура 19), което оказва влияние върху сходимостта на модела при трениране с голям брой ръчно анотирани изображения.

Подобрение може да се направи и в използваната функция на грешката, като в нея се отразят, чрез тегла за отделните обучителни пример, допълнителни фактори, като например – закриване на части от лицето, екстремна осветеност, по-редки изражения. Може да се добави и адаптиране на обучителната стъпка (learning rate) и на минималното тегло което назначаваме на примерите с малка грешка в изчислената ориентация на лицето (минимума от 0.1 във формула 20), спрямо достигнатата епоха на трениране. Целта на това би било в по-късните епохи, когато общата грешка е достигнала ниски нива да се даде повече принос на примерите, които продължават да дават голямо отклонение. Друг подход в тази насока, който не е разгледан в никоя от цитираните тук разработки е да се придадат различни тежести на грешките спрямо различните характерни точки, тъй като, както упоменахме в секция 2.1, позициите на някои от точките се локализират по-надеждно от други, като това е в сила както за машинното им засичане, така и за ръчното им анотиране, прилагано при изготвянето на повечето налични набори от данни. В [36] е описан експеримент при изготвянето на набора от данни 300W, при който част от изображенията са анотирани независимо от трима експерти, след което техните анотации са сравнени (фигура 20).



***Фигура 20**: Отклонения при ръчното анотиране на 68-те точки от набора от данни 300W. Източник: [36]*

Възможно е и да се модифицира архитектурата на мрежата чрез премахване или добавяне на още блокове от тип Inverted residual, както и добавянето на повече от 3 нива конволюции в края на мрежата. Това би имало за цел постигане на още по-висока скорост или на по-висока точност, като едното би било за сметка на другото.

Източници

[1] Akca, Devrim. (2003). Generalized Procrustes Analysis and its applications in Photogrammetry.

[2] Bailenson, Jeremy & Pontikakis, Emmanuel (Manos) & Mauss, Iris & Gross, James & Jabon, Maria & Hutcherson, Cendri & Nass, Clifford & John, Oliver. (2008). Real-time classification of evoked emotions using facial feature tracking and physiological responses. International Journal of Human-Computer Studies. 66. 303-317. 10.1016/j.ijhcs.2007.10.011.

[3] Bulat, Adrian & Tzimiropoulos, Georgios. (2016). Human Pose Estimation via Convolutional Part Heatmap Regression. 9911. 10.1007/978-3-319-46478-7_44.

[4] Bulat, Adrian & Tzimiropoulos, Georgios. (2017). How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a Dataset of 230,000 3D Facial Landmarks). 10.1109/ICCV.2017.116.

[5] Burgos-Artizzu, Xavier & Perona, Pietro & Dollár, Piotr. (2013). Robust Face Landmark Estimation under Occlusion. Proceedings of the IEEE International Conference on Computer Vision. 1513-1520. 10.1109/ICCV.2013.191.

[6] Çeliktutan, Oya & Ulukaya, Sezer & Sankur, Bulent. (2013). A comparative study of face landmarking techniques. EURASIP Journal on Image and Video Processing. 2013. 10.1186/1687-5281-2013-13.

[7] Chateau, Thierry & Duffner, Stefan & Garcia, Christophe & Blanc, Christophe & Naturel, Xavier & Yan, Yongzhe. (2018). A survey of deep facial landmark detection.

[8] Cootes, T.F. & Taylor, Chris & Cooper, D.H. & Graham, Jim. (1995). Active Shape Models-Their Training and Application. Computer Vision and Image Understanding. 61. 38–59. 10.1006/cviu.1995.1004.

[9] Cootes, Timothy & Edwards, Gareth & Taylor, Christopher. (1998). Active Appearance Models. IEEE Trans. Patt. Anal. Mach. Intell.. 23. 484-498.

[10] Cristinacce, David & Cootes, Timothy. (2006). Feature Detection and Tracking with Constrained Local Models. Pattern Recognit.. 41. 929-938. 10.5244/C.20.95.

[11] Dantone, Matthias & Gall, Juergen & Fanelli, Gabriele & Van Gool, Luc. (2012). Real-time facial feature detection using conditional regression forests. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 10.1109/CVPR.2012.6247976.

[12] Dibeklioglu, Hamdi & Salah, Albert & Gevers, T.. (2012). A Statistical Method for 2-D Facial Landmarking. Image Processing, IEEE Transactions on. 21. 844 - 858. 10.1109/TIP.2011.2163162.

[13] Dollár, Piotr & Welinder, Peter & Perona, Pietro. (2010). Cascaded Pose Regression. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 10.1109/CVPR.2010.5540094.

[14] Dong, Xuanyi & Yan, Yan & Ouyang, Wanli & Yang, Yi. (2018). Style Aggregated Network for Facial Landmark Detection. 10.1109/CVPR.2018.00047.

[15] Edwards, G.J. & Taylor, Christopher & Cootes, T.F.. (1998). Interpreting face images using active appearance models. 300 - 305. 10.1109/AFGR.1998.670965.

[16] Felzenszwalb, Pedro & Huttenlocher, Daniel. (2005). Pictorial Structures for Object Recognition. International Journal of Computer Vision. 61. 55-79. 10.1023/B:VISI.0000042934.15159.49.

[17] Feng, Zhen-Hua & Kittler, Josef & Awais, Muhammad & Huber, Patrik & Wu, Xiao-Jun. (2018). Wing Loss for Robust Facial Landmark Localisation with Convolutional Neural Networks. 10.1109/CVPR.2018.00238.

[18] Ghiasi, Golnaz & Fowlkes, Charless. (2015). Occlusion Coherence: Detecting and Localizing Occluded Faces.

[19] Gross, Ralph & Matthews, Iain & Baker, Simon. (2004). Appearance-Based Face Recognition and Light-Fields. IEEE transactions on pattern analysis and machine intelligence. 26. 449-65. 10.1109/TPAMI.2004.1265861.

[20] Gross, Ralph & Matthews, Iain & Baker, Simon. (2005). Generic vs. Person Specific Active Appearance Models. Image and Vision Computing. 23. 1080-1093. 10.1016/j.imavis.2005.07.009.

[21] Guo, Xiaojie & Li, Siyuan & Zhang, Jiawan & Ma, Jiayi & Ma, Lin & Liu, Wei & Ling, Haibin. (2019). PFLD: A Practical Facial Landmark Detector.

[22] Howard, Andrew & Pang, Ruoming & Adam, Hartwig & Le, Quoc & Sandler, Mark & Chen, Bo & Wang, Weijun & Chen, Liang-Chieh & Tan, Mingxing & Chu, Grace & Vasudevan, Vijay & Zhu, Yukun. (2019). Searching for MobileNetV3. 1314-1324. 10.1109/ICCV.2019.00140.

[23] Hu, Changbo & Feris, Rogerio & Turk, Matthew. (2003). Real-time view-based face alignment using active wavelet networks. Proc. IEEE Int'l Workshop Analysis and Modeling of Faces and Gestures. 215 - 221. 10.1109/AMFG.2003.1240846.

[24] Jin, Xin & Tan, Xiaoyang. (2016). Face Alignment In-the-Wild: A Survey. Computer Vision and Image Understanding. 162. 10.1016/j.cviu.2017.08.008.

[25] Kazemi, Vahid & Sullivan, Josephine. (2014). One Millisecond Face Alignment with an Ensemble of Regression Trees. 10.13140/2.1.1212.2243.

[26] Khabarlak, Konstantin & Koriashkina, Larysa. (2020). Fast Facial Landmark Detection and Applications: A Survey. 10.13140/RG.2.2.32735.07847/1.

[27] King, Davis. (2009). Dlib-ml: A Machine Learning Toolkit. Journal of Machine Learning Research. 10. 1755-1758. 10.1145/1577069.1755843.

[28] Köstinger, Martin & Wohlhart, Paul & Roth, Peter M. & Bischof, Horst. (2011). Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization. Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. 2144-2151. 10.1109/ICCVW.2011.6130513.

[29] Kumar, Abhinav & Marks, Tim & Mou, Wenxuan & Wang, Ye & Jones, Michael & Koike-Akino, Toshiaki & Cherian, Anoop & Liu, Xiaoming & Feng, Chen. (2020). LUVLi Face Alignment: Estimating Landmarks' Location, Uncertainty, and Visibility Likelihood. 10.1109/CVPR42600.2020.00826.

[30] Kurakin, Alexey & Goodfellow, Ian & Bengio, Samy. (2016). Adversarial examples in the physical world.

[31] Le, Vuong & Brandt, Jonathan & Lin, Zhe & Bourdev, Lubomir. (2012). Interactive Facial Feature Localization. 10.1007/978-3-642-33712-3_49.

[32] Liu, Rosanne & Lehman, Joel & Molino, Piero & Such, Felipe & Frank, Eric & Sergeev, Alex & Yosinski, Jason. (2018). An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution.

[33] Liu, Yaojie & Jourabloo, Amin & Ren, William & Liu, Xiaoming. (2017). Dense Face Alignment. 1619-1628. 10.1109/ICCVW.2017.190.

[34] Ouanan, Hamid & Ouanan, Mohammed & Aksasse, B.. (2016). Facial landmark localization: Past, present and future. 487-493. 10.1109/CIST.2016.7805097.

[35] Qian, Shengju & Sun, Keqiang & Wu, Wayne & Qian, Chen & Jia, Jiaya. (2019). Aggregation via Separation: Boosting Facial Landmark Detector With Semi-Supervised Style Translation. 10152-10162. 10.1109/ICCV.2019.01025.

[36] Sagonas, Christos & Antonakos, Epameinondas & Tzimiropoulos, Georgios & Zafeiriou, Stefanos & Pantic, Maja. (2016). 300 Faces In-The-Wild Challenge: database and results. Image and Vision Computing. 47. 10.1016/j.imavis.2016.01.002.

[37] Sagonas, Christos & Tzimiropoulos, Georgios & Zafeiriou, Stefanos & Pantic, Maja. (2013). 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. 397-403. 10.1109/ICCVW.2013.59.

[38] Sandler, Mark & Howard, Andrew & Zhu, Menglong & Zhmoginov, Andrey & Chen, Liang-Chieh. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. 4510-4520. 10.1109/CVPR.2018.00474.

[39] Saragih, Jason & Lucey, Simon & Cohn, Jeffrey. (2011). Deformable Model Fitting by Regularized Landmark Mean-Shift. International Journal of Computer Vision. 91. 200-215. 10.1007/s11263-010-0380-4.

[40] Sauer, Patrick & Cootes, Tim & Taylor, Christopher. (2011). Accurate Regression Procedures for Active Appearance Models. BMVC 2011 - Proceedings of the British Machine Vision Conference 2011. 10.5244/C.25.30.

[41] Sun, Yi & Wang, Xiaogang & Tang, Xiaoou. (2013). Deep Convolutional Network Cascade for Facial Point Detection. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 3476-3483. 10.1109/CVPR.2013.446.

[42] Szegedy, Christian & Zaremba, Wojciech & Sutskever, Ilya & Bruna, Joan & Erhan, Dumitru & Goodfellow, Ian & Fergus, Rob. (2013). Intriguing properties of neural networks.

[43] Tong, Yan & Wang, Yang & Zhu, Zhiwei & Ji, Qiang. (2007). Robust facial feature tracking under varying face pose and facial expression. Pattern Recognition. 40. 3195-3208. 10.1016/j.patcog.2007.02.021.

[44] Toshev, Alexander & Szegedy, Christian. (2013). DeepPose: Human Pose Estimation via Deep Neural Networks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 10.1109/CVPR.2014.214.

[45] Tresadern, Phil & Sauer, Patrick & Cootes, Timothy. (2010). Additive Update Predictors in Active Appearance Models. 1-12. 10.5244/C.24.91.

[46] Wang, Xinyao & Bo, Liefeng & Li, Fuxin. (2019). Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression. 6970-6980. 10.1109/ICCV.2019.00707.

[47] Wu, Wenyan & Qian, Chen & Yang, Shuo & Wang, Quan & Cai, Yici & Zhou, Qiang. (2018). Look at Boundary: A Boundary-Aware Face Alignment Algorithm. 2129-2138. 10.1109/CVPR.2018.00227.

[48] Wu, Yue & Ji, Qiang. (2019). Facial Landmark Detection: A Literature Survey. International Journal of Computer Vision. 127. 10.1007/s11263-018-1097-z.

[49] Wu, Yue & Wang, Zuoguan & Ji, Qiang. (2013). Facial Feature Tracking Under Varying Facial Expressions and Face Poses Based on Restricted Boltzmann Machines. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 3452-3459. 10.1109/CVPR.2013.443.

[50] Xiong, Xuehan & De la Torre, Fernando. (2013). Supervised Descent Method and Its Applications to Face Alignment. Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 532-539. 10.1109/CVPR.2013.75.

[51] Yang, Heng & Patras, Ioannis. (2013). Privileged information-based conditional regression forest for facial feature detection. 1-6. 10.1109/FG.2013.6553766.

[52] Zhang, Kaipeng & Zhang, Zhanpeng & Li, Zhifeng & Qiao, Yu. (2016). Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Processing Letters. 23. 10.1109/LSP.2016.2603342.

[53] Zhu, Xiangyu & Lei, Zhen & Liu, Xiaoming & Shi, Hailin & Li, Stan. (2016). Face Alignment Across Large Poses: A 3D Solution. 146-155. 10.1109/CVPR.2016.23.