ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ
ФАКУЛТЕТ ПО ТЕЛЕКОМУНИКАЦИИ
КАТЕДРА „РАДИОКОМУНИКАЦИИ И ВИДЕОТЕХНОЛОГИИ“

маг. инж. Теодора Георгиева Сечкова

ОПРЕДЕЛЯНЕ НА ЕМОЦИОНАЛНИ ИЗРАЖЕНИЯ И ПОВЕДЕНИЕ ОТ ЧОВЕШКИ ЛИЦА НА БАЗАТА НА ИЗОБРАЖЕНИЯ И ВИДЕОПОСЛЕДОВАТЕЛНОСТИ

АВТОРЕФЕРАТ

на дисертация за присъждане на образователна и научна степен

„ДОКТОР”

Област на висше образование: 5. Технически науки

Професионално направление: 5.3 Комуникационна и компютърна техника

Научна специалност: Телевизионна и видеотехника

Научен ръководител: доц. д-р инж. Агата Христова Манолова

София, 2023

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Радиокомуникации и видеотехнологии“ към Факултет по Телекомуникации при Технически Университет – София на редовно заседание, проведено на 09.10.2023 г. (протокол № 36).

Публичната защита на дисертационния труд ще се състои на 04.12.2023 г. (понеделник) от 13:00 часа в Конферентната зала на БИЦ на Технически университет – София на открито заседание на научното жури, определено със заповед ОЖ–5.3-58/16.10.2023г. на Ректора на ТУ-София в състав:

1. Доц. д-р инж. Агата Манолова – ТУ-София, научен ръководител
2. Проф. д-р инж. Снежана Плешкова-Бекярска – ТУ-София
3. Проф. д-р инж. Александър Бекярски - пенсионер
2. Доц. д-р инж. Страхил Соколов – ВУТП
3. Проф. д-р инж. Димитър Димитров – пенсионер

Рецензенти:

1. Проф. д-р инж. Снежана Плешкова-Бекярска
2. Проф. д-р инж. Александър Бекярски – ТУ-София

Резервни членове за журито:

1. Доц. д-р инж. Иво Драганов – ТУ-София
2. Проф. д-р инж. Станимир Садинов – ТУ-Габрово

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет по Телекомуникации, блок 1, стая 1254 и на Интернет страницата на Технически Университет – София.

Дисертантът е редовен докторант към катедра „Радиокомуникации и видеотехнологии“, Факултет по Телекомуникации. Изследванията по дисертационният труд са направени от автора, като резултатите от тях са публикувани. Част от изследванията са вследствие на работата по научноизследователски проекти.

Автор: Маг. инж. Теодора Георгиева Сечкова

Заглавие: Определяне на емоционални изражения и поведение от човешки лица на базата на изображения и видеопоследователности

Тираж: 30 броя

Отпечатано в ИПК на Технически университет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

Изражението на лицето играе главна роля в социалния живот на човека, то е основа на невербалната комуникация, изразяването на емоции и възприемането на човешкото поведение. Погледът и кимането са основни поведенчески сигнали при комуникацията между хората, визуалната информация, възприемана при движението на устните, помага за разбиране на речта, движенията на мускулите на лицето предават чувствата и намеренията. Израженията са внезапни, автоматични, несъзнателни реакции на емоциите изпитвани от човека. Всички хора изразяват емоции на лицата си по един и същи начин, те са универсални и не зависят от раса, етнос, пол, националност, възраст, религия или която и да е друг демографска величина. В продължение на десетилетия изследването на човешкото поведение и в частност на човешките изражения е цел на психологията. С развитието на технологиите, търсенето на нови възможности за подобряване на комуникацията между човек и машина се насочва към създаването на машинни интерфейси моделиращи човешкото поведение. Разпознаването на изражението на лицето и съответната емоция, която то предава, е ключова стъпка в разпознаване на поведенческите сигнали и последващо моделиране на поведението. Последвалият напредък в алгоритмите за машинно обучение и компютърно зрение прави възможно прилагането на разпознаването на емоции в разнообразни области като развлекателната индустрия, медицината и криминалистиката.

Настоящата дисертация се занимава с определянето на изражения и поведение от човешки лица на базата на изображения и видеопоследователности. Анализът на човешките изражения и емоционални реакции намира приложение в системите за препоръка на персонализирано съдържание, както и за анализ на потребителя и целево рекламиране на продукти. В здравеопазването такъв тип системи помагат за ранно откриване на симптоми на депресия, невродегенеративни и психични заболявания. Информацията за емоционалните реакции на учениците се използва за подобряването на учебното съдържание и за оценка на тяхната ангажираност при дистанционно обучение. Не на последно място, системи за разпознаване на израженията на лицето допринасят за откриване и предотвратяване на редица измами.

Цел на дисертационния труд, основни задачи и методи за изследване

Целта на настоящата дисертация е разработване на методи и алгоритми за разпознаване на емоционални изражения от човешки лица на базата на статични изображения и видео последователности.

От поставената цел на дисертационния труд произтичат и следните *основни задачи*:

1. Разработване на алгоритми за разпознаване на основни мускулни движения в изображения на лица чрез автоматично подравняване на ключови точки от лицето и подбор на признаци на базата на графи.

2. Разработване на алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности чрез мрежи базирани на механизъм за внимание и класификация чрез конволюционни невронни мрежи работещи върху графи.

3. Функционална проверка на действието на разработените алгоритми, чрез програмна реализация и стандартни бази данни, за оценка на тяхната ефективност

Методологията на изследванията в дисертацията включва използване на теоретичен и симулационен подход. Теоретичният подход е приложен при разработването на методите и алгоритмите, въз основа на литературата. Симулационният подход е свързан с тестването на алгоритмите, използвайки стандартни бази данни, с цел определяне на тяхната ефективност.

Научна новост

В дисертационният труд са предложени методи и алгоритми за определяне на емоционални изражения от човешки лица на базата на изображения и видеопоследователности. Предложените методи стъпват на съвременни концепции и подходи в машинното обучение и разпознаването и класификацията на изображения. Използвани са съвременни софтуерни библиотеки и бази данни за реализация на алгоритмите.

Практическа приложимост

Представените нови методи и алгоритми в този труд са приложими като компоненти от системи в множество комерсиални и научни области като сигурност и безопасност, поведенчески науки, медицина, комуникации и образование. В областта на сигурността, израженията на лицето играят основна роля в разпознаването на лъжа и оценка на правдивостта на твърденията на даден човек. В областта на медицината и поведенческите науки, лицевите изражения са средство за откриване на ментални процеси като болка, депресия, стрес, умора, безпокойство. В сферата на образованието, израженията на лицето на учениците могат да бъдат източник на информация за учителя за определяне на вниманието, което те отделят на представената им тема.

Публикуване на резултатите от дисертационното изследване

Направените анализи, предложените подходи и получените резултати са представени в общо **5** авторски публикации на престижни *международни конференции* и една в престижно *международно научни списание.* Всички **5** публикации са в съавторство. Налични са общо **13** цитирания, всички от които са в индексирани издания.

Международните конференции са: *2016 IEEE 8th International Conference on Intelligent Systems, 2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting, 2019 IEEE Wireless Communications and Networking Conference Workshop, 2019 European Conference on Networks and Communications.*

Международните научни списания са: *International Journal of Signal Processing,Volume 2017*.

Структура и обем на дисертационния труд

Дисертационният труд е в обем от **110** страници формат А4 и съдържа увод, четири основни глави, заключение с изложени основни приноси, списък на фигурите, списък на таблиците, списък на използваните съкращения, списък на проектите, в които е участвал докторанта, списък с публикациите по дисертацията, списък на използваната литература. Изложението на дисертационния труд, направено в **4** глави, съдържа **27** фигури и **12** таблици. Използвани са **129** литературни източници. Номерата на фигурите, таблиците и математическите изрази в автореферата съответстват на тези в дисертационния труд.

В настоящия дисертационен труд са предложени алгоритми за определяне на емоционални изражения от човешки лица на базата на изображения и видеопоследователности. Първият алгоритъм е предназначен за класификация на основни мускулни движения от статични изображения на лица и следва структурата на конвенционалните методи за разпознаване на емоции. Вторият алгоритъм е предназначен за класификация на спонтанни емоционални изражения от видеопоследователности и е базиран на дълбоки невронни мрежи. Ефективността на предложените алгоритми е изследвана с бази данни, популярни в научната литература.

В първа глава е направен анализ на основните модели на човешките емоции използвани от алгоритмите за разпознаване на емоционални изражения. Отбелязано е различното предназначение и наложилото се приложение на отделните модели за целите на компютърното зрение. Извършен е задълбочен анализ на конвенционалните методи за разпознаване на емоционални изражения, използващи извличане на признаци специално пригодени за конкретна задача и на методите базирани на дълбоки невронни мрежи. Разгледани са в детайли всички етапи на конвенционалните методи и тяхното предназначение. Извършен е сравнителен анализ на алгоритми за разпознаване на емоционални изражения от лица докладвани в литературата, които се базират на конвенционални методи. Разгледани са основните характеристики на дълбоките невронни мрежи, както и на дълбоки невронни мрежи със специализирани архитектури според типа на входните данни. Анализирани са основните различия между дълбоките невронни мрежи и конвенционалните методи. Извършен е сравнителен анализ на алгоритми за разпознаване на емоционални изражения от лица докладвани в литературата, които се базират на дълбоки невронни мрежи. Разгледани са свойствата характерни за видеопоследователностите и разликите им със статичните изображения в задачата за откриване на емоционални изображения. Направен е сравнителен анализ на методите докладвани в литературата за работа с видеопоследователности.

Във втора глава е разработен алгоритъм за разпознаване на основни мускулни движения от изображения на лица чрез подбор на признаци на базата на графи. Алгоритъмът работи в два етапа, етап на обучение и етап на тест, използващи съответно обучителна и тестова извадка. Първата стъпка от алгоритъма е предварителната обработка на входните изображенията, последвана от втора стъпка на извличане на признаци, състояща се от откриване и подравняване на ключови точки от лицето чрез алгоритъма Supervised Descend Method (SDM) и същинското извличане на признаци чрез оператора Scale Invariant Feature Transform (SIFT). В третата стъпка се извършва подбор на признаци на базата на графи чрез алгоритъма Multi-Cluster Feature Selection (MCFS). В последния етап се извършва класификация на

извлечените признаци с алгоритъма Kernel Support Vector Machines (KSVM) с радиални базисни функции.

В трета глава е разработен алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности чрез невронни мрежи с механизъм за внимание и конволюционни мрежи работещи върху графи. Алгоритъмът работи в два етапа, етап на обучение и етап на тест, използващи съответно обучителна и тестова извадка. Алгоритъмът работи в две части. В първата част е използвана пространствено-времева мрежа с механизъм за внимание за извличане на признаци от видеопоследователности. Мрежата се състои от пространствен кодер реализиран чрез мрежа Swin Transformer за извличане на пространствени зависимости и времеви кодер реализиран чрез мрежа с механизъм за внимание за извличане на времеви зависимости от входните данни. Във втората част е използвана граф-конволюционна невронна мрежа за моделиране на зависимостите между отделните емоции и изчисляване на параметрите на класификатор, който се използва за класификация на пространствено-времевите признаци извлечени от входните данни в първата част.

В четвърта глава са извършени експериментални изследвания на разработените алгоритми. Експериментите са проведени върху публично достъпни бази данни съдържащи както изиграни така и спонтанни емоционални реакции. Базата данни използвана за обучение и тест на алгоритъма в глава първа е заснета в контролирани условия, докато базата данни използвана от алгоритъма в глава втора е заснета от самите участници в неконтролирани условия. Експериментите включват сравнителен анализ на няколко конфигурации на алгоритмите и сравнение с резултатите на алгоритми предложени в литературата, които използва същите или сходни условия на оценка.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. Анализ на състоянието на проблема по литературни данни

1.1. Модели на човешките емоции

В основополагащата публикация на Екман и Фрезен [1], се дефинират шест основни емоции, на които отговарят шест основни лицеви изражения, които се разпознават универсално от хората от всички култури. Това са щастие (радост), тъга, страх, отвращение, гняв и изненада, които се допълват в литературата и от съставни емоции, образувани от няколко основни. Детайлно описание на израженията на лицето задава системата Facial Action Coding System (FACS) [2], която описва лицевите изражения като комбинация от 44 основни мускулни движения, наречени Action Units (AU). За разлика от дискретните модели категоризиращи отделни емоции, моделът „валентност-възбуда” [3] дава по-гъвкава дефиниция на емоциите като точки от едно непрекъснато пространство, вместо да ги ограничава в краен брой дискретни състояния.

1.2. Анализ на методите за разпознаване на емоционални изражения от изображения на лица и видео последователности

Първоначално, използваните методи базиращи се на извличане на признаци и последваща класификация, т.нар. конвенционални методи, биват подобрени от методи базирани на дълбоки невронни мрежи, а в последствие добри резултати постигат и методите базирани на т.нар. механизъм за внимание. Типично за конвенционалните методи е предварителният избор на признаци, за разлика от методите базирани на невронни мрежи, където алгоритъмът сам открива интересни признаци във входните изображения, разчитайки на наличието на голям обем данни за обучение. Методите базирани на механизъм за внимание са друг тип дълбоки невронни мрежи, който доби популярност през последните три години. След първоначалния им успех в областта на обработката на естествен език, те биват адаптирани успешно за целите на компютърното зрение. В сравнение с предходните две групи методи, тези базирани на механизъм за внимание имат най-малко предварителни знания за входните данни, което води до нуждата от огромни по обем бази данни за обучението им.

Основните стъпки в един конвенционален метод са предварителна обработка, извличане на признаци и класификация. Предварителната обработка на входните изображения е необходима първа стъпка при всички видове алгоритми в областта на компютърното зрение. Най-често тя цели премахване на ненужна информация от изображенията, подобряване на качеството им, намаляване на обема от информация. Извличането на признаци е втората основна стъпка при конвенционалните методи за разпознаване на емоции. Целта ѝ е извличането на полезна информация от изображенията, която да улесни следващия етап на класификация, чрез максимизиране на между-класовата вариация, и минимизиране на вътре-класовата такава. Конвенционалните методи разчитат на предварителното подбиране на подходящи признаци, според типа на данните и задачата, което прави тази стъпка от решаваща важност за крайния резултат на алгоритъма. Методите за извличане на признаци от изображения на лица биват разделяни на две групи, геометрични и методи базирани на външност. Чрез стъпката за извличане на признаци изображението се представя в ново пространство на признаците, улесняващо последващата задача на класификация, като от това не винаги следва намаляване на размерността на входните данни. При една част от методите, признаковият вектор е с много по-голяма размерност от размерността на извадката, което води до т.нар. „проклятие на размерността”. За избягването му се използват методи, които биха могли да се разделят на две групи, според типа на преобразуването на данните, линейни и нелинейни статистически методи, както и нелинейни методи използващи многообразия. Класификацията е финалната стъпка на методите за разпознаване на емоционални изражения, при която класификаторът категоризира израженията в един от предварително зададени класове. В зависимост от начина, по който е поставена задачата, класовете могат да отговарят на основните емоции, на едно или комбинация от основни мускулни движения.

За разграничаване между конвенционалните методи и т.нар. „дълбоки” методи за машинно обучение, тяхната архитектура се разглежда като поредица от йерархични слоеве на изчисление. С понятието „дълбочина” на архитектурата се означава броя на изчислителните слоеве. Използвайки тази терминология, конвенционалните методи могат да бъдат представени условно като съставени от два слоя, слой на извличане на признаци или представяне на входните данни в ново пространство подходящо за класификация и слой

осъществяващ самата класификация. Сравнявайки се с представените по този начин конвенционални методи, невронните мрежи биват наричани дълбоки, когато се състоят от три или повече слоя на изчисление, въпреки че няма формално правило задаващо това разделение. Основните предимства на дълбоките невронни мрежи са свързани с големия брой слоеве. Увеличаването на броя слоеве, през които преминават входните данни, дава възможност за последователно извличане на признаци, използвайки на входа на всеки следващ слой изходът от предходния. От друга страна, това позволява извличането на йерархични признаци, което води до построяване на различни нива на абстракция, което се свързва с устойчивост към локални промени във входните данни. Въпреки, че теоретичните предимства на дълбоките невронни мрежи са отдавна известни, тяхното практическо приложение е забавено от трудности при обучението им. Няколко взаимно свързани фактора в последствие допринасят за успешното им развитие и прилагане [4], [5]:

- *Увеличен капацитет*. Основната идея зад дълбоките невронни мрежи е връзката между по-добрата генерализация и увеличения брой параметри на мрежата, най-вече чрез добавяне на повече слоеве към мрежата. За целта са необходими достатъчна изчислителна мощ, за осъществяване на обучението и наличието на достатъчно голяма база данни.

- *Нарастваща изчислителна способност на хардуера*. Обучението на съвременните невронни мрежи е тежка изчислителна задача, изискваща оптимизацията на милиони параметри. Специализиран софтуер и напредъкът в развитието на графичните процесори са основните фактори движещи успеха в областта.

- Наличието на големи бази данни. Дълбоките невронни мрежи се нуждаят от много голям размер входни данни, за успешно извличане на описателни признаци. Изследвания в областта разкриват логаритмична зависимост между успешното справяне със задачи на компютърното зрение и размера на обучителната извадка.

Когато е известно, че входните данни притежават определена пространствена или времева структура, какъвто е случаят с изображенията и видеопоследователностите, дълбоките невронни мрежи се видоизменят, създавайки специализирани архитектури, които да се възползват от тази структура: конволюционни невронни мрежи, рекурентни невронни мрежи, невронни мрежи с механизъм за внимание и други.

Входните данни могат да бъдат разглеждани като два различни типа, статични, когато разглеждаме отделни изображения и динамични, когато разглеждаме не само отделните статични изображения или кадри от видеопоследователности, а и взимаме предвид времевите зависимости между тях. Във втория случай, една видеопоследователност е разглеждана цялостно като поредица от кадри в рамките на един времеви диапазон и като отделен обект на класификацията. Чрез моделиране на времевите вариации се постига по-добро представяне на слабо изразени емоционални изражения, подобрява се разпознаването на изражения близки по между си като движения на лицето. Чрез използването на видеопоследователности се повишава резултата от работата на алгоритъма при лоши условия като ниска осветеност, ниска разделителна способност или промяна в осветеността и позицията на главата. Тези предимства идват за сметка на повишената изчислителна

сложност на алгоритмите, поради увеличения размер на входните данни, които те трябва да обработят.

1.3. Дефиниране на целта и основните задачи на настоящата дисертация

Целта на настоящата дисертация е разработване на методи и алгоритми за разпознаване на емоционални изражения от човешки лица на базата на статични изображения и видео последователности.

От поставената цел на дисертационния труд произтичат и следните *основни задачи*:

1. Разработване на алгоритми за разпознаване на основни мускулни движения в изображения на лица чрез автоматично подравняване на ключови точки от лицето и подбор на признаци на базата на графи.

2. Разработване на алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности чрез мрежи базирани на механизъм за внимание и класификация чрез конволюционни невронни мрежи работещи върху графи.

3. Функционална проверка на действието на разработените алгоритми, чрез програмна реализация и стандартни бази данни, за оценка на тяхната ефективност.

ГЛАВА 2. Разработване на алгоритъм за разпознаване на емоционални изражения от изображения на лица

Голяма част от алгоритмите за разпознаване на емоции от изображения на лица, изхождат от трудовете на Екман, в които той представя идеята, че основните емоции имат съответстващи прототипни изражения. Предложените алгоритми целят да разпознаят именно тези шест изражения. Въпреки това, в естествена среда израженията рядко съответстват с точност на тези прототипи. Често емоциите се предават чрез слабо изразени промени в чертите на лицето, напр. повдигане на веждите при изпитване на изненада. Именно за да обхванат тези промени и да опишат обективно разнообразните и сложни изражения на лицето, друга част от алгоритмите залагат на разпознаването на отделни или комбинации от основни мускулни движения базирани на FACS. Поради своята изчерпателност, FACS дава възможности за откриване на нови комбинации от черти на лицето, отговарящи на съответни емоционални състояния [6], [7]

В глава втора е предложен алгоритъм за разпознаване на основни мускулни движения от изображения на лица, който е базиран на автоматично откриване и подравняване на ключови точки от лицето и извличане на устойчиви признаци от околността на тези точки. Предимството на предложения алгоритъм, спрямо другите в литературата е използването на алгоритъм за точно откриване на координатите на ключови точки в комбинация с извличане на признаци, устойчиви на ротация, мащабиране и промени в осветлението. Това води до извличане на признаци от точните координати на ключови елементи от човешкото лице за изразяването на емоции - очи, уста, нос и т.н. Друго предимство на алгоритъма е подборът на признаци базиран на методите за вграждане на графи, които са част от нелинейните методи с многообразия, запазващи геометричните свойства на данните, което води до подбор на

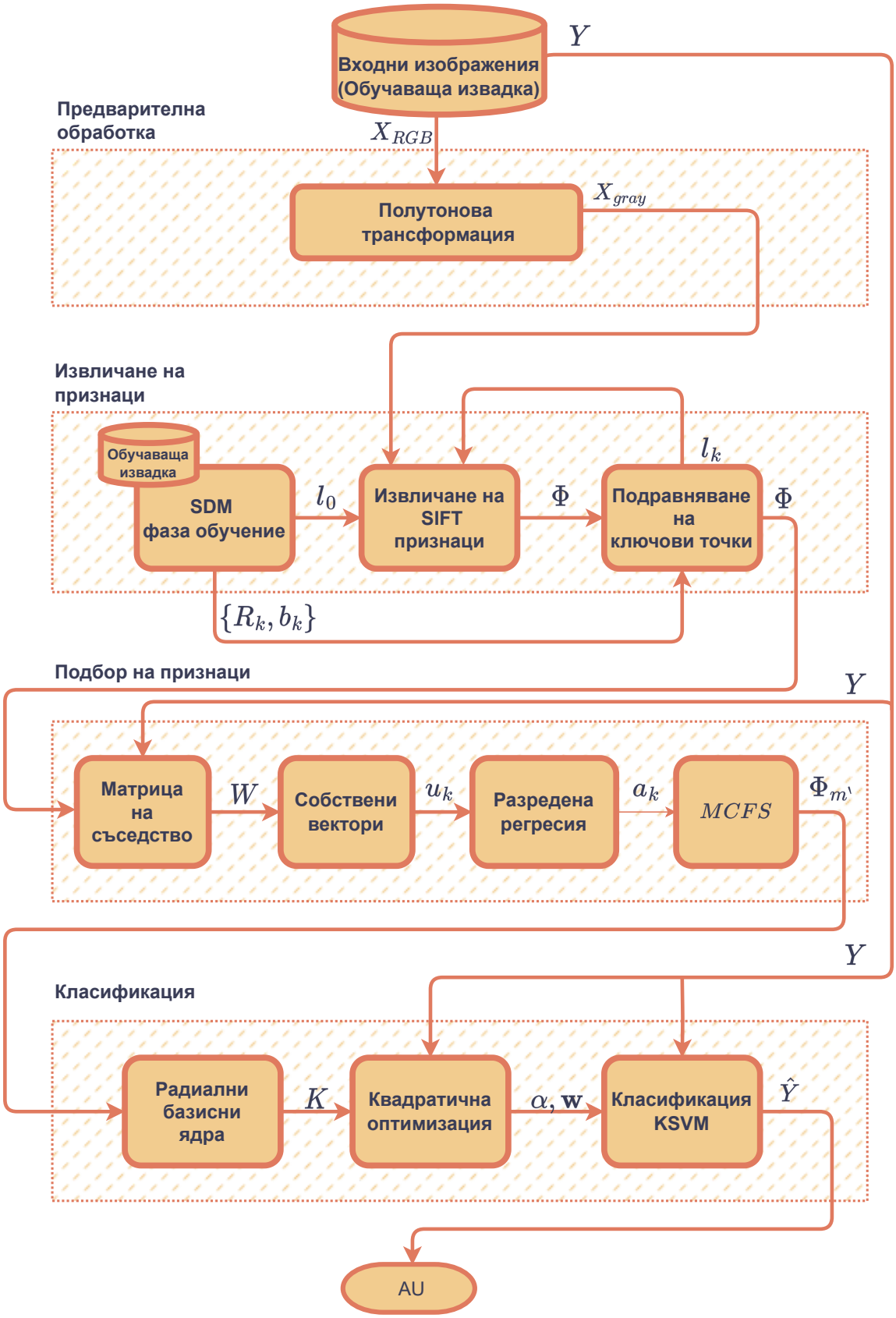
информативни признаци и подобрява резултатите на алгоритъма за класификация. Алгоритъмът работи в два етапа, етап на обучение и тестови етап.

2.1. Общо описание на алгоритъм за разпознаване на основни мускулни движения от изображения на лица чрез подбор на признаци на базата на графи

Разпознаването на всяко основно мускулно движение се дефинира като класификация на входното изображение в два класа, положителен клас, съдържащ съответния AU и отрицателен клас. Алгоритъмът преминава през етапи на предварителна обработка, извличане и подбор на признаци и класификация. Една част от методите използвани от алгоритъма се базират на обучение, от което следва, че параметрите на всеки един от тях се изчисляват с обучаваща извадка. Поради това, алгоритъмът се реализира в две фази, на обучение и на тест.

Блок-схема на потока на информацията във фазата обучение е представен на Фигура 2.1. Първият етап е предварителна обработка. В този етап се подготвя изображението за извличането на признаци, преобразувайки го в полутоново, за да се намали обема на информация, тъй като цвета не е от съществено значение при разпознаването на изражения. Така обработени, изображенията X_{gray} се предават на етапа за извличане на признаци. В тази стъпка се извличат признаци представляващи геометрията на лицето, която е от основно значение за разпознаването на изражения. Описанието на лицето е под формата на ключови точки разположени на характерни части от лицето, като очи, уста, нос и т.н. При промяна на изражението на лицето, респективно геометрията му, тези точки променят координатите си. Алгоритъмът който се използва за откриване и подравняване на ключовите точки е SDM [8]. Той работи в последователни стъпки, като открива общо 49 ключови точки, представени чрез координатите на изображението. Стъпките от алгоритъма SDM се прилагат върху признаци извлечени чрез оператора SIFT от области на изображението с размери 32×32 пиксела, всяка една от които центрирана върху ключова точка. Алгоритъмът за подравняване на ключови точки преминава през отделен етап на обучение, с друга обучителна извадка, предназначена за откриване на ключови точки. За начални стойности на алгоритъма се задават ключови точки l_0, които представляват средни стойности от обучаващата извадка. Предимството на използването на алгоритъма SDM за извличане на геометрични признаци се състои в неговата устойчивост на промени в позата и осветлението както и на изчислителната му ефективност.

Размерността на извлечените SIFT признаци Ф е много по-голяма от размера на обучаващата извадка, което предполага проява на ефекта на „проклятието на размерността”. За да се избегне този проблем, размерността се намалява чрез етап на избор на признаци. Този етап се базира на анализ на графи и вграждането на признаците в многообразие. В началото на алгоритъма е необходимо построяването на матрица на съседство на графа, W, чрез която се представят приликите между всеки два признакови вектора, които представляват и върхове на графа. За намиране на векторното представяне на върховете на графа, на базата на матрицата на съседство се построява матрицата на Лаплас на графа, L, чрез която се намират собствените вектори u_k, и съответните собствени стойности λ_k. На базата на собствените вектори се формулира задача на регресия, при която се търси разредено подмножество на признаците най-добре описващи вграждането на многообразието.



Фигура 2.1: Блок схема на предложения алгоритъм в глава втора, в етап на обучение.

В следващия етап се използва алгоритъмът Multi-Cluster Feature Selection (MCFS) [9], за избор на $m' < m$ на брой признаци, резултатът, от което е намаляване на размерността. Обучението на нелинейния класификатор SVM се състои в намиране на опорни вектори и построяване на разделяща хиперравнина на базата на извлечените признаци и анотациите от обучителната извадка.

След етапа на обучение, използвайки вече изчислените параметри на отделните алгоритми, се преминава към тестови етап. По време на този етап, на базата на отделна тестова извадка, се определя ефективността на обучения алгоритъм. Стъпките на предварителна обработка и извличане на SIFT признаци не се различават спрямо фазата на обучение, като за откриването на ключови точки се използват коефициентите получени чрез обучението. При стъпката избор на признаци се използват индексите на избраните признаци, изчислени във фазата на обучение, за да се изберат m' на брой признака. Избраните признаци се подават към SVM класификатора, който използва опорните вектори изчислени през фазата на обучение, за да определи класа основно мускулно движение, $\hat{Y}$.

2.2. Откриване на ключови точки и извличане на признаци със SDM и SIFT

Локалните признаци извлечени от изображения на лица, са предпочитани пред глобалните поради тяхната устойчивост на промяна на позицията на главата или при закриване на част от лицето [10]. Операторът SIFT е един такъв локален дескриптор, който е инвариантен към промени в мащаба, ориентацията на изображението както и промени в осветлението. Поради тези си желани свойства, SIFT е често използван в алгоритмите за разпознаване на изражения на лицето [11], [12]. SIFT работи в няколко стъпки, като първата от тях е намирането на екстремни точки от изображението, които в последващите стъпки се използват за ключови точки, от околността, от които се съставя признаковият вектор. Алгоритъмът намира екстремните точки чрез декомпозиция на изображението на базата на Difference of Gaussians (DoG).

Разчитайки на SIFT сам да открие ключови точки в изображението, води до това, че те не съвпадат с точките и зоните, които се свързват с изразяването на емоции (вежди, очи, нос, уста и т.н). Освен, че това влошава резултатите при разпознаването на изражения, броят на ключовите точки, откривани от алгоритъма, не може да бъде предварително определен и варира между две отделни изображения, дори и на близки изражения, което затруднява извличането на признаков вектор, подходящ за използване от алгоритми за класификация като SVM [13], [14]. Взимайки под внимание тези проблеми, редица алгоритми използват предварително дефинирани точки, описващи изображението на лицето и прилагат оператора SIFT върху тях [15], [16].

От една страна, ключовите точки променят координатите си при промяна на геометрията на лицето, в резултат от съответните мускулни движения при формирането на изражението, което води до нуждата от използване на алгоритми за подравняване на точките на лицето. От друга, операторът SIFT е недиференцируем, което в случая на използването му за извличане на признаци от ключови точки, затруднява задачата на подравняването. За решаване на такава задача се изисква използването на оптимизационни методи от втори ред, необходимо условие, за което е оптимизираната функция да бъде аналитично диференцируема, а когато то не е изпълнено се прилагат числови апроксимации, които са с висока изчислителна

сложност. Точно тези проблеми адресира SDM, който е метод за минимизиране на нелинейни функции.

SDM обединява дискриминтивните и параметричните методи, но за разлика от параметричните методи, не съставя модел на формата и външността от обучителната извадка, а работи директно с координатите на ключови точки от изображението. Това води до по-добра генерализация при работа в реални условия. Основно предимство на SDM пред дискриминативните методи, е способността му да решава нелинейни задачи на най-малките квадрати в общ вид, както и да превъзхожда по бързодействие методите използващи регресия само с една стъпка на изчисление. Освен това алгоритъмът удовлетворява целта на задачата за постигане на устойчивост при промяна на позата и осветеността на изображенията.

В сравнение с общата задача за разпознаване на обекти, операторът SIFT има своите недостатъци при прилагането му върху изображения на лица. Изображенията на лица имат по-малко високо-контрастни структури и контури. Тъй като ключови точки с нисък контраст или намиращи се по дължината на контур, могат да бъдат премахнати при подбора на ключови точки в оригиналната имплементация на оператора SIFT, това може да доведе до премахване на интересни ключови точки от лицето. За да се избегне този проблем е необходимо внимателно да се подберат праговете управляващи процеса на избор на ключови точки. Това е процедура на опит и грешка, която води до неоптимални резултати в разпознаването. Друг подход е разделянето на лицето на зони и търсенето на съответствие на признаците в локалната зона, но това води до намаляване на устойчивостта към промени в осветеността. Чрез използването на предварително подбрани ключови точки, които се подравняват за всяко изображение на лицето чрез алгоритъма SDM, се елиминират проблемите типични за изображения на лица и нуждата от намиране на оптимален праг за определяне на ключовите точки.

2.3. Подбор на признаци на базата на графи

В практиката, броят на признаците често може да надмине многократно големината на обучителната извадка, поради трудността при съставянето и анотирането на данни. Това води до ефекта на „проклятието на размерността”, който значително влошава резултатите от класификацията. Такъв е и случаят с избраните в предложения алгоритъм SIFT признаци. Например, избраният брой ключови точки е 49, за всяка от които се построява хистограма със 128 стълба, тогава получената размерност на признаците е 49×128 = 6272, докато една база данни от изражения на лицето, може да бъде в пъти по-малка от този размер. В този случай „проклятието на размерността” е особено изразено и ефективността на класификатора намалява значително. За да се избегнат последиците от голямата размерност на признаковия вектор, е необходимо прилагането на стратегия за намаляване на размерността. Основните подходи в такъв случай са два. Първият е чрез намаляване на размерността, при който се търси линейна или нелинейна комбинация от признаци. Вторият е избор на признаци, при който се избира малка част от най-представителните признаци, спрямо задачата, която се решава.

Предложеният метод в тази работа е алгоритъм за подбор на признаците основан на базата на методите за вграждане на графи [17]. Тези методи са част от обучението чрез

многообразия, доказало се успешно за моделиране на геометричните свойства на данните и прилагано в различни алгоритми за проектиране в подпространство, подбор на признаци и регуляризация [18]. Целта на директното вграждане на граф е да представи всеки връх на графа като вектор от ниска размерност, който запазва приликите между двойките върхове, където приликите се представят чрез матрица на теглата, която характеризира определени статистически или геометрични свойства на входните данни. Векторното представяне на върховете на графа може да бъде получено чрез собствените вектори съответстващи на първите собствени стойности на матрицата на Лаплас на графа при определени ограничения.

Целта е намирането на такова представяне на данните, което да отговаря на многообразие от ниска размерност вградено в първоначалното пространство на данните от висока такава. Това може да се постигне чрез вграждането на граф, който описва връзките между съседни точки от данните, и използване на съответствието на матрицата на Лаплас на графа и операторът на Лаплас-Белтрами на многообразието. Резултатът е ефективен алгоритъм за нелинейно намаляване на размерността, който запазва локалните връзки между данните и също води до тяхната клъстеризация [19]. Алгоритъмът се основава на ролята на оператора на Лаплас-Белтрами за намиране на оптималното вграждане на многообразието. Многообразието е апроксимирано чрез граф построен на базата на съседството между точките, представляващи признаковите вектори, докато операторът на Лаплас-Белтрами се апроксимира чрез матрицата на Лаплас на графа [20]. . Вграждането в подпространство с ниска размерност получаваме като търсим собствените вектори на матрицата на Лаплас на построения граф.

2.4. Класификация на признаци чрез SVM и радиални базисни функции

Класификацията на съответните основни мускулни движения се осъществява чрез алгоритъма KSVM. Входни данни за алгоритъма са признаците подбрани по метода MCFS, а в резултат от класификацията, признаковият вектор отговарящ на входното изображение се причислява към съответния клас, отговарящ на едно основно мускулно движение.

В общия случай, задачите за класификация при реални данни са по-сложни и класовете не могат да бъдат разделени с линейни функции. За разделянето им трябва да бъдат въведени нелинейни такива. Данните могат да се представят в пространство от по-висока размерност, където те са линейно разделими, чрез заместването $\langle \phi_i, \phi_j \rangle \rightarrow \langle g(\phi_i), g(\phi_j) \rangle$. Преминаването към пространство от по-висока размерност би направило изчисленията тежки и би увеличило изчислителната сложност на алгоритъма, но тъй като данните не се използват директно в пространството с висока размерност, използва се само тяхното скаларно произведение, може да бъде приложен метода на ядрото. Чрез заместването на скаларното произведение на данните с ядро съставено от радиални базисни функции, може да се постигне линейно разделяне на данните в пространството с висока размерност, без на практика да се извършват изчисленията в него.

2.5. Изводи от работата по глава втора

- Използването на SDM за откриване и проследяване на ключови точки от лицето, води до предварително известен и дефиниран брой точки, съвпадащ с точките свързващи се с изразяване на емоции, което подобрява резултатите от класификацията, а броят на

ключовите точки между две отделни изображения е предварително известен, което улеснява извличането на признаков вектор, подходящ за използване от алгоритми за класификация.

- Чрез подравняване на координатите на ключовите точки посредством SDM се избягва тяхното изместване при промяна на геометрията на лицето, в резултат от съответните мускулни движения при формирането на изражението.
- Работата на SDM в две отделни фази на обучение и тест, води до неговото бързодействие, без загуба на точност и позволява използването му в реално време при тест фазата.
- Подбор на признаците е необходим, тъй като извлечените признаци са с висока размерност спрямо големината на обучителната извадка.
- Моделирането на признаците чрез вграждане на граф позволява нелинейно намаляване на размерността, което запазва локалните връзки между данните и също води до тяхната клъстеризация.

ГЛАВА 3. Разработване на алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности

Емоционалните изражения на лицето не са статични, те се формират динамично във времето. Всеки израз на емоция преминава през начална фаза, пик и фаза на затихване. Анализирайки видепоследователностите като поредица от статични изображения, се губи времевата зависимост между последователните кадри, от която могат да бъдат извлечени интересни за разпознаването на изражения признаци.

Емоциите възникващи като реакция на външен или вътрешен стимул не са независими една от друга. Често един стимул предизвиква комбинация от емоции и съответстващите им изражения. От своя страна някои емоции възникват едновременно по-често от други, например „изненада” и „щастие” или „гняв” и „тъга”. При разработването на алгоритми за разпознаване на емоционални изражения, моделирането на връзките между отделните емоции носи допълнителна информация и подобрява резултата от разпознаването.

В трета глава е предложен алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности. Алгоритъмът работи в два етапа, етап на обучение и етап на тест и се състои от две части. Първата част представлява извличане на признаци от видеопоследователности на базата на невронна мрежа с механизъм за внимание, която се състои от две под-мрежи, пространствена, извличаща признаци от отделните статични кадри, и времева, целяща да открие зависимостите между кадрите.

Във втората част алгоритъмът цели да моделира зависимостите между отделните емоции, като за целта построява насочен граф от етикетите в базата данни. Всеки етикет съответства на една емоция, която е представена като връх на графа, а реброто между тях отразява зависимостта между двете емоции. От така построения граф, чрез конволюционна невронна мрежа работеща върху графи, Graph Convolutional Network (GCN), се съставят класификатори, които се използват за класификация на извлечените признаци от видеопоследователностите в първата част на алгоритъма.

Предимствата на предложения алгоритъм са извличането на признаци, които комбинират пространствените и времевите характеристики на една емоционална реакция, чрез което се постига устойчивост към частично закриване на лицето, не-фронтална поза или движение на главата. Извличането на признаци става на базата на мрежи с механизъм за внимание, които постигат най-добри резултати в класификацията на изображения и в частност емоции към настоящия момент. Друго предимство на алгоритъма е построяването на класификатори взимащи предвид зависимостите между емоциите.

3.1. Общо описание на алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности чрез невронни мрежи с механизъм за внимание и конволюционни мрежи работещи върху графи

Предложеният метод за разпознаване на спонтанни емоционални изражения от видеопоследователности се състои от две дълбоки невронни мрежи, една пространствено-времева невронна мрежа базирана на механизъм за внимание, за представяне на динамични изражения на лицето и втора конволюционна невронна мрежа работеща с графи, GCN, за моделиране на зависимостите между емоциите.

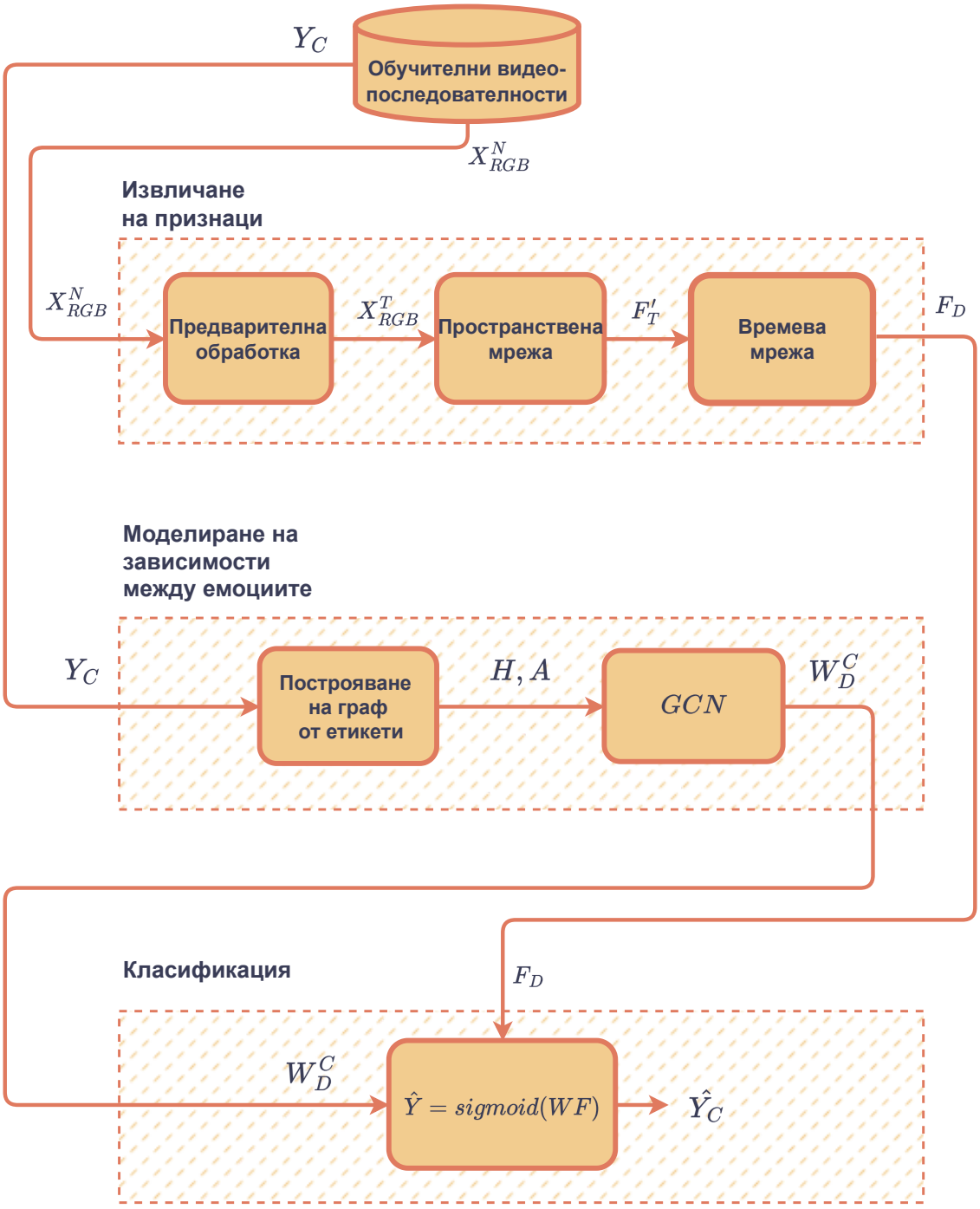
От пространствена гледна точка, цялото изображение на лицето от един кадър се разделя на последователни блокове с еднакъв размер, които могат да бъдат разгледани като поредица от визуални думи. От времева гледна точка, всеки кадър от видеопоследователността може да бъде разгледан като визуална дума. Чрез това представяне на входните данни, механизмът на внимание, по време на етапа на обучение, може да изчисли корелациите между локални признаци от изображението на лицето, както и корелации между времевите признаци, постигайки вградена способност да не се влияе от частично закриване на лицето, не-фронтална поза или движение на главата. Блок схема на алгоритъма от етапа на обучение е представена на Фигура 3.1.

Предложената архитектура се състои от две отделни под-мрежи базирани на механизъм за внимание т.нар. пространствена мрежа и времева мрежа. Пространствената мрежа е изградена на базата на архитектурата Swin Transformer [21], представляваща невронна мрежа с модифициран механизъм на внимание, работещ с отместени прозорци. Предимствата на използването на тази архитектура за пространствена мрежа е нейната гъвкавост спрямо мащаба и линейна изчислителна сложност спрямо размера на изображението. Времевата мрежа се състои от M на брой времеви кодера, които имат за цел извличане на признаци от последователните кадри на лицето чрез използване на времеви контекст.

След предварителна обработка и избор на T на брой кадри от оригиналните видеопоследователности, X_{RGB}^{T}, те се подават на входа на пространствената мрежа, която обработва последователно всеки кадър и извлича T на брой пространствени признакови вектора, F_T', докато последователността от пространствени признаци на съответстващи на кадрите се подава на входа на времевата мрежа, която разглежда цялата последователност като един образец и построява финалния признаков вектор, F_D, с размерност D.

За разпознаването и класификацията в множество класове е важно да се състави ефективен алгоритъм, който обхваща корелациите между отделните емоции и използва тези корелации

за подобряване работата на класификатора. Построяването на граф за моделиране на вътрешните зависимости между етикетите е гъвкав метод за описване на топологичната структура на пространството на етикетите. Всеки етикет от обучителната извадка, който представлява дума описваща съответната емоция, се представя като връх на графа чрез вграждането му в пространство на думите. Чрез използването на GCN, се намира директното съответствие между векторите, върхове на графа, и множество от взаимосвързани класификатори, които могат да бъдат директно приложени върху признаците извлечени от едно изображение, за разпознаването му.



Фигура 3.1: Блок схема на предложения алгоритъм в глава трета, в етап на обучение

Параметрите изчислени през етапа на обучение са споделени между всички класове, в резултат на което класификаторите запазват семантичната структура на пространството на вграждането на думите, където семантично близки думи се намират по-близо едни до други. В същото време градиентите на всеки класификатор могат да повлияят на функцията на построяване на класификатора, което моделира зависимостите между етикетите. Построяването на корелационна матрица на базата на едновременното присъствие на два или повече етикета в една видеопоследователност, позволява на GCN директно да моделира зависимостите между етикетите, чрез което актуализацията на признаците на един връх включва информация от върхове, с които има голяма корелация.

Във втората част от предложения метод, на базата на етикетите на съответните C на брой емоции, Y_C, и техните зависимости, се построява граф. Всяка емоция е представена като връх от графа, а ребрата представляват зависимостите между тях. Чрез конволюционна невронна мрежа работеща с графи, GCN, на базата на графа от етикетите, чрез обучение, се построява множество от C на брой класификатори с размерност C×D, W_D^C.

В последната стъпка на входа на тези класификатори се подават признаковите вектори извлечени от видеопоследователностите. Изходът от класификаторите са множество етикети, $\widehat{Y_C}$, за всяка последователност, присъждащи едновременно всяка последователност към няколко класа с различна интензивност.

По време на тестовия етап, се запазват стъпките в първата част на алгоритъма, като параметрите на мрежите с механизъм за внимание са вече известни след етапа на обучение. Във втората част на алгоритъма, не се налага построяването на граф, а се използва директно вече изчислената матрица W, за класификация на признаците в различните класове.

3.2. Пространствено-времева мрежа с механизъм за внимание

Първа част от предложения метод представлява извличане на информативни признакови вектори от израженията на лицето. Това се постига чрез подаването на входните изображения на невронна мрежа базирана на механизъм за внимание. По конкретно, алгоритъмът се базира на архитектурата на Former-DFER [22], която се състои от две отделни под-мрежи базирани на механизъм за внимание т.нар. конволюционно-пространствена мрежа и времева мрежа. В предложения алгоритъм, конволюционно-пространствената мрежа е заменена с модела Swin Transformer, представляващ невронна мрежа с модифициран механизъм на внимание, работещ с отместени прозорци. Предимствата на използването на тази архитектура за пространствена мрежа е нейната гъвкавост спрямо мащаба и линейна изчислителна сложност спрямо размера на изображението. Времевата мрежа се състои от M на брой времеви кодера, които имат за цел извличане на признаци от изображенията на лицето чрез използване на времеви контекст.

Мрежата Swin Transformer е изградена от слоеве с модифициран механизъм за внимание. Нейната архитектура е разделена на т.нар. „етапи”, които са поредица от слоеве обработващи данните. Първият етап се състои от линейна проекция, последвана от L на брой слоеве с механизъм за внимание. Първият етап запазва размерността на входните блокове. За получаване на йерархично представяне на входното изображение, всеки следващ етап след първия, започва със слой на обединяване на блоковете, който съединява всяка група от съседни блокове с размер 2 × 2. Този слой намалява броя на блоковете два пъти във всяко

направление, еквивалентно на намаляване на резолюцията два пъти, а общият брой блокове става N/4. Следват още слоеве с механизъм за внимание. Общо етапите от този тип са три, като след всеки следващ етап, резолюцията намалява два пъти, чрез което се постига йерархично извличане на признаци от изображението, с признакови вектори близки до тези на конволюционните невронни мрежи. Слоевете с модифициран механизъм за внимание са построени на базата на стандартният механизъм за внимание, но видоизменяйки го за работа с отместени прозорци. Един слой се състои от модифициран механизъм за внимание, последван от MLP с Gaussian Error Linear Unit (GELU) [33] нелинеен оператор. Преди всяка операция има нормиращ слой, а след всяка се добавя и остатъчна връзка.

Механизмът за внимание с множество паралелни клонове, Multi-head Self Attention (MSA), в своя оригинален вид, както и в предложените адаптации за класификации на изображения, работи глобално върху цялото изображение като изчислява зависимостта между един блок с всички останали. Това глобално изчисление има квадратична сложност спрямо броя блокове, на които е разделено изображението, повишавайки значително изчислителната сложност на повечето задачи на компютърното зрение. За подобряване ефективността на алгоритъма, слоевете на Swin мрежата изчисляват вниманието в рамките на локални околности или прозорци на работа. Прозорците разделят блоковете от изображението на равномерни незастъпващи се части. Сложността на стандартния механизъм за внимание е в квадратична зависимост от броя на блоковете, докато при фиксиран брой на блоковете в един прозорец, сложността спрямо блоковете в изображението става линейна. Следователно модифициран механизъм за внимание може да бъде прилаган и при работа с по-голям брой блокове и по-висока резолюция.

Недостатък на предложеният до тук модифициран механизъм за внимание е липсата на връзка между отделните прозорци, което ограничава възможностите за извличане на полезна информация. С цел добавянето на връзка между прозорците, без загуба на ефективност на алгоритъма, слоевете с модифициран механизъм за внимание са разпределени по двойки, с отместени прозорци между всяка една двойка. Първият слой използва регулярна схема за разделяне на изображението като първият прозорец започва от горният ляв ъгъл на изображението. В следващият слой прозорците са отместени от предходния с блока като по този начин се осъществява припокриване с част от предходния слой и се осигурява връзка по между им.

Времевата мрежа се състои от M на брой времеви кодера, като от своя страна всеки кодер се състои от редуващи се слоеве на механизъм за внимание с множество паралелни клонове (multi-head), последван от напълно свързан слой.

3.3. Граф-конволюционна невронна мрежа

Етикетите от бази данни показващи комбинация от едновременно изразяване на няколко емоции в една и съща видеопоследоватлност, могат да послужат за откриване на конкретни зависимости между емоциите. За моделиране на зависимостите между етикетите, съответно на зависимостите между едновременно изразяваните емоции, се построява насочен граф на базата на етикетите. Всяка една от седем емоции е представена като връх на графа, ребрата показват връзките между отделните емоции.

Мрежите от тип GCN работят чрез разпространение на информация между отделните върхове, на базата на корелационна матрица, което прави построяването ѝ съществена част от изграждането на мрежата. В предложения метод, корелационната матрица се построява на базата на данните от обучителната извадка. Корелацията между етикетите в обучителната извадка се определя на базата на тяхното едновременно съществуване в една и съща видеопоследователност. Основната идея зад GCN е първоначалното признаково представяне на един връх да бъде актуализирано чрез разпространяване на информация между върховете. За разлика от стандартната конволюция, която се изпълнява върху локалната структура на изображението в евклидово пространство, целта на GCN е чрез етап на обучение да изчисли параметрите на функцията $f(.,.)$ върху граф G, която има за вход признаковите вектори $H \in \mathbb{R}^{C\times D}$ и съответната корелационна матрица на графа A.

3.4. Класификатор за разпознаване на множество етикети

В конкретната задача за разпознаване на множество етикети, всеки връх на графа съответства на един етикет, а крайният изход от всеки връх е класификатор за съответния етикет, чийто параметри са изчислени по време на етапа на обучение. Матрицата на корелация се построява от съответните етикети в базата данни. Цялата мрежа е съставена от два последователни GCN слоя.

В етапа на обучение се изчисляват взаимосвързани класификатори с параметри матриците $W = \{W_i\}_{i=1}^{C}$. Изчисляването на W през етапа на обучение става на базата на вграждането на етикетите в признаково пространство и функцията изпълнявана от GCN, където с C е отбелязан броят на класовете. Използват се последователни GCN слоеве, където всеки слой l взима на входа си признаковия вектор H^l, докато на изхода му се получава новия вектор H^{l+1}. Изходът от последния слой е матрицата $W \in \mathbb{R}^{C\times D}$, където D е размерността на признаковите вектори, F, извлечени от видеопоследователностите. Чрез прилагане на изчислените класификатори върху съответните признаци получаваме предсказаните класове $\hat{Y} = sigmoid(WF)$.

3.5. Изводи от работата по глава трета

- Чрез извличане на признаци, които комбинират пространствените и времевите характеристики на една емоционална реакция, се постига устойчивост към частично закриване на лицето, не-фронтална поза или движение на главата.
- Използването на Swin Transformer за пространствена мрежа осигурява устойчивост спрямо промени в мащаба на изображението.
- Използването на Swin Transformer подобрява изчислителна сложност на алгоритъма спрямо същият използващ мрежи с глобален механизъм за внимание.
- GCN са ефективен метод, който обхваща корелациите между отделните емоции и подобрява работата на класификатора.

ГЛАВА 4. Експериментални резултати от разработените алгоритми

4.1. Експериментални изследвания на алгоритмите към Глава 2

4.1.1. Експериментални резултати от алгоритъм за класификация на основни мускулни движения от изображения на лица

Експериментите са проведени върху базата данни Extended Cohn-Kanade (CK+) [24]. Базата данни се състои от видеопоследователности съдържащи емоционалните изражения на 210 възрастни индивида, заснети фронтално, в контролирана среда чрез две камери. Демографското разпределение на хората от извадката е следното. Участниците са между 18 и 50 години, 69% жени, 81% евро-американци, 13% афро-американци и 6% от други групи. Участниците са инструктирани да изпълнят последователности от 23 изражения на лицето, които съдържат от едно до няколко основни мускулни движения. Всеки запис започва и свършва с неутрално изражение на лицето, преминавайки през върхова фаза на изражението. Видеопоследователностите са записани в размер 640×480 пиксела, с 8-битови полутонови стойности. От тези записи е съставена база данни, съдържаща общо 593 видеопоследователности с продължителност между 10 и 60 кадъра, започващи с неутрален кадър и завършващи във върховата част на изражението на съответната емоция. Кадрите на върховата част са напълно анотирани според FACS.

За целта на проведените експерименти са използвани всички 593 напълно анотирани кадъра, съдържащи върховата част на изражението. Важно е да се отбележи, че всеки кадър може да съдържа както само едно основно мускулно движение, така и комбинация от няколко движения. Поради спецификите на изразяването на емоции на човешкото лице, по-често срещан е вторият вариант. На базата на тези особености, задачата за класификация се построява, като се формира отделен класификатор в два класа за всяко основно мускулно движение. Всички изображения съдържащи определено основно мускулно движение, само по себе си или като комбинация с други, се включват в положителния клас, а всички останали в отрицателния. По този начин задачата за класификация представлява задача "един към много" и показва колко точно предложеният алгоритъм може да различи едно движение спрямо останалите.

На базата на анотираните основни мускулни движения, базата данни задава и таблица на съответствие на комбинациите от основните мускулни движения и съответната основна емоция, която авторите създават основавайки се на оригиналната таблица на съответствие зададена във FACS ръководството на Екман.

Базата данни в оригиналния си вид е небалансирана, което ще рече, че броят изображения не е равен за отделните класове основни мускулни движения, а за някои от тях е дори твърде малък. За постигане на по-балансирана база данни и за избягване на грешки в разпознаването, спрямо по-рядко представени основни мускулни движения, е подбрана част от базата данни, за която има анотирани поне 50 изображения към клас. Избраните основни мускулни движения както и броя анотирани изображения са представени в Таблица 4.3.

Таблица 4.3: Използваните AU и броя изображения, в които присъстват.

AU	Наименование	Брой	AU	Наименование	Брой
1	Вътрешно повдигане на вежда	173	9	Сбръчкване на нос	74
2	Външно повдигане на вежда	116	12	Опъване на ъгъл на устна	111
4	Снижаване на вежда	191	15	Издърпване надолу на ъгъл на устна	89
5	Повдигане на горни клепачи	102	17	Повдигане на брадичка	196
6	Повдигане на бузи	122	25	Разделяне на устни	287
7	Присвиване на клепачи	119	27	Разтягане на уста	81

Друго наблюдение върху така получената база данни, е че отрицателният клас съдържа много по-голям брой изображения от положителния, което може да доведе до изместване на резултата на класификатора към отрицателния клас. За да се избегне този проблем, броят на изображенията от отрицателния клас са намалени, така че съотношението изображения от отрицателния клас към положителния да не надминава два пъти.

Оценката на алгоритъма се извършва чрез процедурата leave-one-person-out (LOPO), където изображенията на израженията от един индивид се използват за тест, а останалите за обучение. Чрез тази процедура се избягва зависимост на обучението към отделен индивид.

Признаковите вектори, извлечени с оператора SIFT, са съставени чрез построяване на хистограма със 128 стълба от локален прозорец с размер 32×32, центриран върху всяка една от откритите чрез SDM 49 ключови точки от лицето. Хистограмите са конкатенирани в един признаков вектор с размерност $49 \times 128 = 6272$. Изборът на признаци се извършва върху този признаков вектор в два варианта на алгоритъма MCFS, с обучение със и без учител. И в двата случая броят на запазените признаци е определен експериментално, чрез изпълнение на алгоритъма при различен брой признаци. В случая на обучение без учител, тегловият коефициент на прилика σ е избран експериментално чрез крос-валидация със стъпка 5. На базата на получените резултати, е избран брой на признаците 350 за обучението без учител и 400 за обучението с учител. Класификацията се извършва чрез нелинеен SVM класификатор с ядро на базата на радиални базисни функции. Всеки от SVM параметрите е избран експериментално чрез крос-валидация със стъпка 5. Пълен списък на параметрите на алгоритъма е представен в Таблица 4.4.

Таблица 4.4: Параметри на алгоритъма за разпознаване на основни мускулни движения на базата на графи.

Параметър	Стойност
Размер на изображенията след детектора на лица	48 x 48 пиксела
Брой на ключови точки	49
Брой на избраните признаци	350/400 елемента
Параметър на гаусовото ядро на MCFS σ_1	0.1 (експериментално)
Регуляризиращ параметър на MCFS β	2 (експериментално)
Параметър на ядрото на KSVM σ_2	0.5 (експериментално)

Поради това, че се използва процедурата LOPO, се изчисляват средната стойност (Mean) и стандартното отклонение (STD) на коефициента на разпознаване, разгледан като случайна величина. Средната стойност на разпознаване на алгоритъма е сравнена с алгоритъма предложен в [25]. Сравнението включва и резултатите от прилагане на алгоритъма в двата случая, със и без учител. При изборът на алгоритми за сравнение, е трудно да се намери експеримент със сходна постановка, използващ същата база данни и начин на оценка. В случая, целта на избора е алгоритми за класификация на основни мускулни движения от изображения, извършващи експериментите си върху базата данни CK+, използващи процедурата LOPO. Такава е работата на [33], с разликата, че авторите комбинират две бази данни и използват всички налични основни мускулни движения. Друга работа с близка постановка е [26], които използват само 5 основни мускулни движения и крос-валидация за оценка на алгоритъма, което прави невъзможно директното сравнение. Средната стойност на коефициента на разпознаване на избраните за сравнение алгоритми, са представени в Таблица 4.7.

Таблица 4.7: Резултати от експериментите за определяне на ефективност на предложения алгоритъм и сравнението му с други.

Алгоритъм	Разпознаване %
Valstar и Pantic [25]	90.2
Предложения алгоритъм (без учител)	90.1
Предложения алгоритъм (с учител)	92.7

4.1.2. Анализ на резултатите

Резултатите от сравнителният анализ показват, че предложеният алгоритъм е по-ефективен от други докладвани в литературата, при сходни или еднакви условия на тест. Очаквано, алгоритъмът постига по-високи резултати при работата си с учител 92.7%, но дори и при работата без учител, резултатът от класификацията е съпоставим с докладваните, 90.1% .На база на тези резултати може да се направи извод, че избора на признаци чрез SDM и SIFT е ефективен, и избраните признаци чрез MCFS са информативни. Също така, влиянието на „проклятието на размерността" е намалено, което се доказва чрез постигнатото намаляване на размерността от 20 пъти.

4.2. Експериментални изследвания на алгоритмите към Глава 3

4.2.1. Експериментални резултати от алгоритъм за разпознаване на емоционални реакции от видеопоследователности.

Експериментите са проведени върху базата данни Hume-Reaction [27], която се състои от видеопоследователности заснемащи човешки реакции към визуални емоционални стимули и анотирани спрямо интензитета на емоционалните изражения в седем категории. Базата данни се състои от повече от 70 часа видео записи на 2222-ма участници от САЩ (1138) и Южна Африка (1084) на възраст между 18 и 49 години. Участниците реагират спонтанно на голям брой емоционални стимули. Един образец от базата данни представлява видеопоследователност анотирана от самите участници за интензитета на 7 емоционални

реакции по скалата от 1 до 100. Реакциите използвани за съставянето на базата данни са обожание, забавление, притеснение, отвращение, емпатия към болка, страх и изненада. Видеопоследователностите са записани чрез собствените уеб камери на участниците в обстановка по техен избор включваща различни по вид фонов шум и осветление. Участниците записват своите спонтанни емоционални изражения с различна степен на интензивност.

Емоционалните реакции на участниците са провокирани от емоционални стимули в специално създадена за целта база данни [28]. За целта, авторите на проучването събират 2185 кратки видеоклипа, с дължина 5 секунди, заснемащи широк спектър от емоционални ситуации. Видеоклиповете са събрани чрез търсене в публични бази данни и уеб търсачки и изобразяват ситуации значими от психологична гледна точка, включващи раждания и бебета, сватби, страдание или смърт, паяци и змии, умилителни животни, изкуство и архитектура, природни пейзажи, природни бедствия, война, носталгични филми, храна, танци, рискови каскади и много други.

Базата данни е разделена на обучителна, тестова извадка и извадка за верификация в съотношение 60% − 20% − 20%. При разделението са взети предвид дължината, участниците и отбелязаните реакции. Всяка видеопоследователност съдържа една единствена реакцията на един стимул. Анотациите са нормализирани в интервала [0 : 1]. Броят на участниците и съотношението на отделните извадки са показани в Таблица 4.8.

Таблица 4.8: Съотношението на отделните извадки в базата данни Hume-Reaction.

Извадка	Брой участници	Времетраене
Обучителна	1334	51:04:02
Верификация	444	14:59:27
Тест	444	14:48:21
Общо	2222	74:26:19

Авторите на базата данни предоставят необработените видеопоследователности. С цел уеднаквяване на броя на кадрите и работата с входни данни с еднакъв размер, всяка видеопоследователност се разделя първо на 8 сегмента, от всеки сегмент на случаен принцип се избират 2 кадъра, получената дължина на така формираните входни последователности е 16 кадъра. Всеки кадър се преоразмерява с размер 112×112 пиксела. За откриване на лица в изображенията се използва предварително обучен модел Multi-task Cascaded Convolutional Networks (MTCNN) [29].

Алгоритъмът е изцяло реализиран на Python чрез библиотеката PyTorch [30]. За обучението на моделите е използвана оптимизация Stochastic Gradient Descent (SGD) с коефициент на обучение 0.05 и степен на намаляване 0.5 с период 5 епохи, размерът на един batch е 16, а общата продължителност на обучението е 60 епохи. За функция на загубите е използвана Mean Squared Error (MSE).

Таблица 4.9: Параметри на етап на обучение.

Параметър	Стойност
Метод за оптимизация	SGD
Коефициент на обучение	0.05
Степен на намаляване	0.5
Период на намаляване	5 епохи
Обща продължителност на обучението	60 епохи

Параметрите на пространствено-времевата мрежа са както следва. За пространствена мрежа е използван моделът Swin-T преминал през предварително обучение. Времевата мрежа се състои от 3 времеви кодера, размерът на входните данни за една видеопоследователност е T=16 кадъра, размерът на признаковия вектор е 512.

Параметрите на конволюционната мрежа работеща върху графи са както следва. За извличане на векторно представяне на етикетите е използван модела 300-dim Glove [31] обучен върху базата данни от Wikipedia. Параметрите за построяване на корелационната матрица са τ = 0.14 и p = 0.25. GCN мрежата се състои от два слоя, размера на изходните данни е 512.

Таблица 4.10: Параметри на алгоритъма за разпознаване на емоционални реакции от видеопоследователности..

Параметър	Стойност
Размер на един кадър	112 x 112 пиксела
Брой кадри	16
Размер на признаков вектор	512 елемента
Брой слоеве времеви кодери	3
Брой слоеве GCN	2
Праг на корелационна матрица τ_1	0.14 (експериментално)
Нормализиращ коефициент на корелационна матрица p	0.25 (експериментално)

За оценка на алгоритъма е използван осреднен коефициент на корелация на Pearson (ρ), който е използван за оценка на линейната корелация между предсказаните емоционални реакции и целевите стойности. Резултатите от проведените експерименти са показани в Таблица 4.11. Предложеният метод е представен в два варианта, с използването на GCN и без и е сравнен с базовия резултат върху базата данни Hume-Reaction получен чрез BiLSTM невронна мрежа и признаци извлечени чрез VGGFace2 [32] както и с резултатите публикувани в [33], който е избран заради използването на същата база данни и оценка на алгоритъма.

Резултатите за различните класове са показани в Таблица 4.12. Най-висок резултат е постигнат за класа „забавление”, а най-нисък за „отвращение”.

Таблица 4.11: Резултати от експериментите за определяне на ефективност на предложения алгоритъм и сравнението му с други.

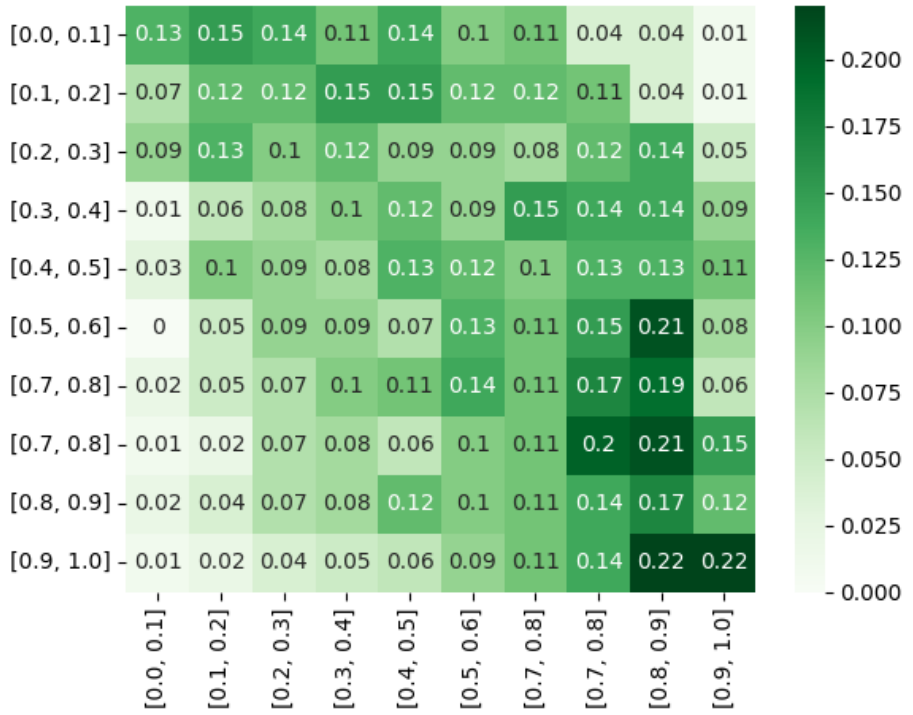
Алгоритъм	ρ
BiLSTM + VGGFace2 [32]	0.2488
ViPER (само видео) [33]	0.2712
ViPER (аудио и видео) [33]	0.3025
Предложения алгоритъм (без GCN)	0.3321
Предложения алгоритъм (с GCN)	0.3375

Таблица 4.12: Резултати от разпознаването на различните класове.

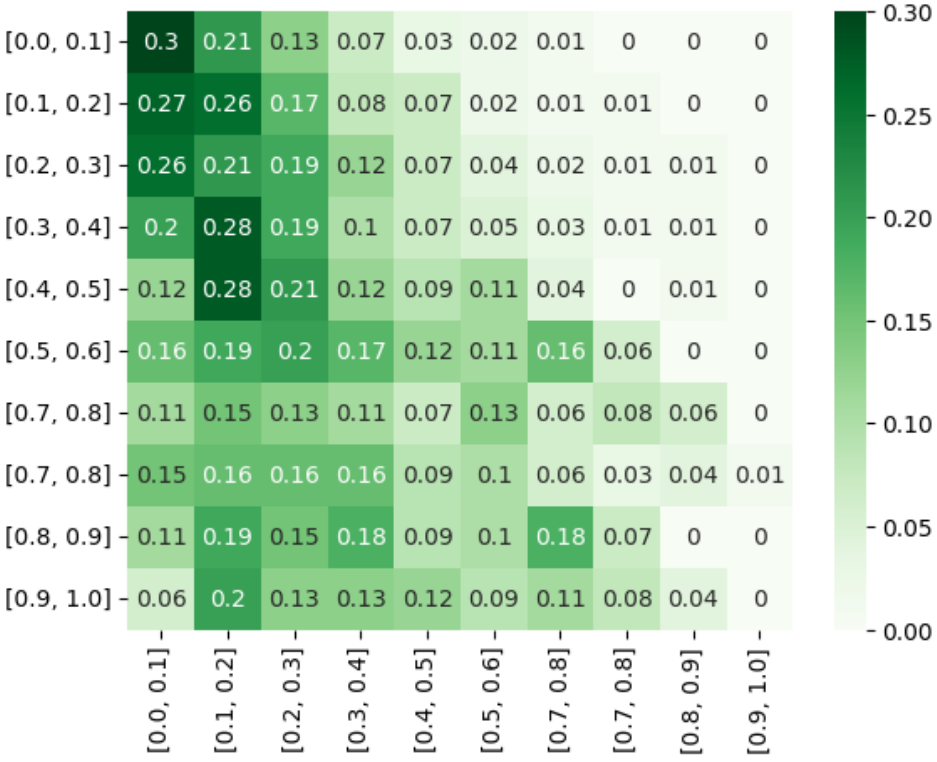
Емоция	ρ
Обожание	0.3281
Забавление	0.3668
Притеснение	0.3501
Отвращение	0.297
Емпатия към болка	0.319
Страх	0.359
Изненада	0.3405

4.2.2. Анализ на резултатите

Резултатите от направения анализ показват, че предложеният алгоритъм е ефективен при разпознаването на емоционални реакции от видеопоследователности и надминава с голям процент базовия резултат върху използваната база данни както и други алгоритми от литературата, като дори надминава резултатите на алгоритъм използващ и аудио данни за подобряване на класификацията. Алгоритъмът постига по-високи резултати при работата си с GCN. Фигури 4.8. и 4.9. показват резултатите от направения допълнителен анализ на двата класа забавление" и „отвращение". Построени са матрици на объркване от изчислените стойности за интензивността на емоциите и стойностите към резултатите от анотациите на базата данни. Забелязва се, че класът „забавление" е по-добре представен около главния диагонал, а по-добри резултати се получават при стойности на висока интензивност на изразяване, докато стойностите за класа „отвращение" са разпределени в интервала [0, 0.3]. Тази разлика би могла да бъде отдадена на разпределението на етикетите в обучителната извадка. Друго обяснение би могло да бъде способността на класификатора да научава по-лесно емоцията „забавление", поради специфичните изражения на лицето. На база на тези резултати може да се направи извода, че извличането на признаци чрез пространствено-времева мрежа е ефективно, а използването на класификатори построени на база зависимостите между емоциите с GCN допринася за по-висок резултат на класификация.



Фигура 4.8: Матрица на грешката за клас „забавление”.



Фигура 4.9: Матрица на грешката за клас „отвращение”.

III. ПРИНОСИ В ДИСЕРТАЦИОННИЯ ТРУД

Научно-приложните приноси са:

1. Разработен е алгоритъм за разпознаване на емоционални изражения от изображения на лица на базата на основни мускулни движения, комбиниращ Supervised Descent Method и оператора SIFT за постигане на работа в реално време и устойчивост към промени в околната среда. Последващият подбор на признаци на базата на вграждане на графи подобрява дискриминативността и избягва ефекта на „проклятието на размерността". (Фигура 2.1, Фигура 2.2)

2. Разработен е алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности чрез невронни мрежи с механизъм за внимание. Използването на две подмрежи с механизъм за внимание позволява едновременно извличане на пространствени и времеви зависимости. В допълнение са използвани конволюционни мрежи работещи върху графи за моделиране на корелациите между емоциите и подобряване на точността на разпознаването. (Фигура 3.1, Фигура 3.2)

Приложните приноси са:

1. Програмна реализация и експериментални изследвания на алгоритъм за подравняване на ключови точки на лицето, извличане и подбор на признаци на базата на анализ на графи. (Фигура 4.2, Фигура 4.3, Таблица 4.5, Таблица 4.6, Таблица 4.7)

2. Програмна реализация и експериментални изследвания на алгоритъм за класификация на основни мускулни движения чрез KSVM и радиални базисни функции. (Фигура 4.4, Таблица 4.5, Таблица 4.6, Таблица 4.7)

3. Програмна реализация и експериментални изследвания на алгоритъм за разпознаване на спонтанни емоционални изражения от видеопоследователности. (Фигура 4.8, Фигура 4.9, Таблица 4.11, Таблица 4.12)
4. Програмна реализация и експериментални изследвания на алгоритъм моделиране на зависимости между емоциите чрез граф. (Фигура 4.7, Таблица 4.11, Таблица 4.12)

IV. ЗАКЛЮЧЕНИЕ

Разработените в настоящата дисертация алгоритми за определяне на емоционални изражения и поведение от човешки лица на базата на изображения и видеопоследователности демонстрират подобрени резултати в сравнение с публикации в същата област към момента на разработването им и показват, че могат да намерят приложение в редица области подобряващи комуникацията между човек и компютър. Резултатите от алгоритъма за разпознаване на основни мускулни движения, построен чрез извличане на признаци специално пригодени за конкретна задача, демонстрират че конвенционалните методи имат добра ефективност при малък обем от данни за обучение. При работа с големи бази данни, използването на дълбоки невронни мрежи, по-конкретно комбинация от различни видове мрежи с механизъм за внимание, за извличане на максимално информативни признаци от изображения и видеопоследователности, води до успешно разпознаване на емоционални реакции и тяхната интензивност. Методите за предварително обучение и фина настройка позволяват използването на дълбоки невронни мрежи вече преминали през етап на обучение върху различна база данни, което значително ускорява етапа на обучение.

V. ИЗПОЛЗВАНИ СЪКРАЩЕНИЯ

AU	Action Unit
CNN	Convolutional Neural Network
DNN	Deep Neural Network
DoG	Difference of Gaussians
FACS	Facial Action Coding System
GCN	Graph Convolutional Network
KSVM	Kernel SVM
LOPO	Leave-One-Person-Out
MCFS	Multi-Cluster Feature Selection
MLP	Multilayer Perceptron
MSA	Multi-head Self Attention
MSE	Mean Square Error
SDM	Supervised Descent Method
SIFT	Scale Invariant Features Transform
SVM	Support Vector Machines

VI. СПИСЪК С ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИЯТА

[D1] **T. Sechkova,** K. Tonchev and A. Manolova, "Action Unit recognition in still images using graph-based feature selection,"2016 IEEE 8th International Conference on Intelligent Systems (IS), Sofia, Bulgaria, 2016, pp. 646-650, doi: 10.1109/IS.2016.7737496.

[D2] **K. Tonchev**, N. Neshov, A. Manolova, V. Poulkov, "Expression recognition using sparse selection of log-Gabor facial features.", In 2017 *Fourth International Conference on Mathematics and Computers in Sciences and in Industry (MCSI)*, (pp. 11-15). IEEE.

[D3] **T. Sechkova**, M. Paolino, and D. Raho. "Virtualized infrastructure managers for edge computing: Openvim and openstack comparison."In 2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), pp. 1-6. IEEE, 2018**.**

[D4] **T. Sechkova**, E. Barberis, and M. Paolino. "Cloud and edge trusted virtualized infrastructure manager (vim)-security and trust in openstack."In 2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW), pp. 1-6. IEEE, 2019.

[D5] **T. Sechkova**, E. Barberis, and M. Paolino. "Secure location-aware VM deployment on the edge through OpenStack and ARM TrustZone."In 2019 European Conference on Networks and Communications (EuCNC), pp. 278-282. IEEE, 2019.

VII. ИЗПОЛЗВАНА ЛИТЕРАТУРА

[1] P. Ekman и W. V. Friesen, „Constants across cultures in the face and emotion.," Journal of personality and social psychology, т. 17, № 2, с. 124, 1971.

[2] P. Ekman и W. V. Friesen, „Facial action coding system," Environmental Psychology & Nonverbal Behavior, 1978.

[3] J. A. Russell, „A circumplex model of affect.," Journal of personality and social psychology, т. 39, № 6, с. 1161, 1980.

[4] J. Heaton, „Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning: The MIT Press, 2016, 800 pp, ISBN: 0262035618," Genetic programming and evolvable machines, т. 19, № 1-2, с. 305—307, 2018.

[5] C. Sun, A. Shrivastava, S. Singh и A. Gupta, „Revisiting unreasonable effectiveness of data in deep learning era," в Proceedings of the IEEE international conference on computer vision, 2017, с. 843—852.

[6] J. Lien, T. Kanade, J. Cohn и C.-C. Li, „Automated facial expression recognition based on FACS action units," в Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, с. 390—395.

[7] M. Pantic и L. Rothkrantz, „Facial action recognition for facial expression analysis from static face images," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), т. 34, № 3, с. 1449—1461, 2004..

[8] X. Xiong и F. De la Torre, „Supervised descent method and its applications to face alignment," в Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, с. 532—539.

[9] D. Cai, C. Zhang и X. He, „Unsupervised feature selection for multi-cluster data," в Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, 2010, с. 333—342

[10] K. Yu, Z. Wang, M. Hagenbuchner и D. D. Feng, „Spectral embedding based facial expression recognition with multiple features," Neurocomputing, т. 129, с. 136—145, 2014.

[11] Y. Tian, T. Kanade, and J. F. Cohn, „Facial expression recognition", in Handbook of face recognition, Springer, 2011, pp. 487–519.

[12] J Kriˇzaj, V Struc и N Paveˇsi´c, „Adaptation of SIFT features for face recognitionˇ under varying illumination," в The 33rd International Convention MIPRO, IEEE, 2010, с. 691—694.

[13] S. Berretti, B. Ben Amor, M. Daoudi и A. Del Bimbo, „3D facial expression recognition using SIFT descriptors of automatically detected keypoints," The Visual Computer, т. 27, с. 1021—1036, 2011.

[14] H. Soyel и H. Demirel, „Localized discriminative scale invariant feature transform based facial expression recognition,“ Computers & Electrical Engineering, т. 38, № 5, с. 1299—1309, 2012.

[15] F. Ren и Z. Huang, „Facial expression recognition based on AAM–SIFT and adaptive regional weighting,“ IEEJ Transactions on Electrical and Electronic Engineering, т. 10, № 6, с. 713—722, 2015

[16] W. Zheng, H. Tang, Z. Lin и T. S. Huang, „A novel approach to expression recognition from non-frontal face images,“ в 2009 IEEE 12th international conference on computer vision, IEEE, 2009, с. 1901—1908..

[17] J. Wang, L. Yin, X. Wei и Y. Sun, „3D facial expression recognition based on primitive surface feature distribution,“ в 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), IEEE, т. 2, 2006, с. 1399—1406

[18] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang и S. Lin, „Graph embedding and extensions: A general framework for dimensionality reduction,“ IEEE transactions on pattern analysis and machine intelligence, т. 29, № 1, с. 40— 51, 2006.

[19] Y. Fu и Y. Ma, Graph embedding for pattern analysis. Springer Science & Business Media, 2012.

[20] M. Belkin и P. Niyogi, „Laplacian eigenmaps for dimensionality reduction and data representation,“ Neural computation, т. 15, № 6, с. 1373—1396, 2003.

[21] M. Belkin и P. Niyogi, „Towards a theoretical foundation for Laplacian-based manifold methods,“ Journal of Computer and System Sciences, т. 74, № 8, с. 1289—1308, 2008.

[22] Z. Liu, Y. Lin, Y. Cao и др., „Swin transformer: Hierarchical vision transformer using shifted windows,“ в Proceedings of the IEEE/CVF international conference on computer vision, 2021, с. 10012—10022.

[23] Z. Zhao и Q. Liu, „Former-dfer: Dynamic facial expression recognition transformer,“ в Proceedings of the 29th ACM International Conference on Multimedia, 2021, с. 1553—1561.

[24] D. Hendrycks и K. Gimpel, „Gaussian error linear units (gelus),“ arXiv preprint arXiv:1606.08415, 2016.

[25] K. Zhang, Z. Zhang, Z. Li и Y. Qiao, „Joint face detection and alignment using multitask cascaded convolutional networks,“ IEEE signal processing letters, т. 23, № 10, с. 1499—1503, 2016..

[26] A. Paszke, S. Gross, F. Massa и др., „Pytorch: An imperative style, highperformance deep learning library,“ Advances in neural information processing systems, т. 32, 2019.

[27] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar и I. Matthews, „The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression,“ в 2010 ieee computer society conference on computer vision and pattern recognition-workshops, IEEE, 2010, с. 94—101.

[28] L. Christ, S. Amiriparian, A. Baird и др., „The muse 2022 multimodal sentiment analysis challenge: humor, emotional reactions, and stress,“ в Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, 2022, с. 5—14.

[29] A. S. Cowen и D. Keltner, „Self-report captures 27 distinct categories of emotion bridged by continuous gradients,“ Proceedings of the national academy of sciences, т. 114, № 38, E7900—E7909, 2017.

[30] M. Valstar и M. Pantic, „Fully Automatic Facial Action Unit Detection and Temporal Analysis,“ в 2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW’06), 2006, с. 149—149.

[31] J. Pennington, R. Socher и C. D. Manning, „Glove: Global vectors for word representation,“ в Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, с. 1532—1543

[32] M. H. Mahoor, M. Zhou, K. L. Veon, S. M. Mavadati и J. F. Cohn, „Facial action unit recognition with sparse representation,“ в 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG), IEEE, 2011, с. 336—342.

[33] [123] L. Vaiani, M. La Quatra, L. Cagliero и P. Garza, „Viper: Video-based perceiver for emotion recognition,“ в Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge, 2022, с. 67—73.

TECHNICAL UNIVERSITY OF SOFIA
FACULTY OF TELECOMMUNICATIONS
DEPARTMENT "COMMUNICATION NETWORKS"

Teodora Georgieva Sechkova, M.Sc.

FACIAL EXPRESSIONS AND BEHAVIOR RECOGNITION FROM STILL IMAGES AND VIDEO SEQUENCES

ABSTRACT of Ph.D. THESIS

The main topic of this thesis is facial expressions and behavior recognition from still images and video sequences capturing human faces. Facial expressions are the foundation of non-verbal communication between humans and play a main role in displaying human emotions. Automatic analysis of facial expressions is applied in content recommendation systems, for early disease detection in healthcare, in fraud detection or analysis of students' reactions and improving educational materials. The main goals of the thesis are determined after state-of-the-art analysis and have two main directions, developing algorithms for Action Unit (AU) recognition in still images with handcrafted feature extraction and dynamic facial expression recognition with deep neural networks.

The proposed algorithm for AU recognition from still images is based on the conventional computer vision approaches and uses handcrafted features. It is a supervised machine learning algorithm. The key step of the algorithm is facial key points detection and alignment with SDM in combination with SIFT feature extraction which leads to facial features extracted from the exact facial landmarks (eyes, nose, mouth) invariant to changes in scale, rotation, illumination. The second part of the algorithm is feature selection based on graph embedding and sparse regression based on MCFS. Sparse feature selection is used as a dimensionality reduction technique to avoid the "curse of dimensionality" but also has the benefit of providing descriptive features preserving the geometry of the input data space.

The second algorithm proposed in this thesis is an algorithm for dynamic facial expression recognition based on spatio-temporal transformer network and graph-convolutional network (GCN). It is a supervised machine learning algorithm. The algorithm uses a spatio-temporal transformer network to extract feature representations from video frames capturing dynamic facial expressions. The spatial part of the network is based on Swin Transformer for capturing spatial correlations, the temporal part is an MSA-based coder for capturing temporal correlation between frames. GCN is used for modeling the dependences between emotions in terms of dataset labels and learning a classifier for the extracted spatio-temporal features.

To study their performance, the algorithms are implemented in software and their performance is evaluated using popular datasets containing facial expression images and videos. For objective evaluation of the algorithms, their performance is compared to others, proposed in the scientific literature.