ТЕХНИЧЕСКИ УНИВЕРСИТЕТ – СОФИЯ
Факултет по Телекомуникации
Катедра "Радиокомуникации и Видеотехнологии"

Маг. инж. Пламен Атанасов Христов

ВИЗУАЛЕН АНАЛИЗ НА ПОВЕДЕНИЕТО НА ИНДИВИДИ В КИБЕРФИЗИЧНИ СИСТЕМИ

А В Т О Р Е Ф Е Р А Т

на дисертация за придобиване на образователна и научна степен
"ДОКТОР"

Професионално направление: 5.3. Комуникационна и компютърна техника
Научна специалност: Телевизионна и видеотехника

Научни ръководители: Проф. д-р инж. Огнян Любенов Бумбаров
Доц. д-р инж. Агата Христова Манолова

СОФИЯ, 2024 г.

Дисертационният труд е обсъден и насочен за защита от Катедрения съвет на катедра „Радиокомуникации и видеотехнологии“ към Факултет по Телекомуникации на ТУ-София на редовно заседание, проведено на 28.10.2024 г..

Публичната защита на дисертационния труд ще се състои на 28.01.2025 г. (вторник) от 15:00 часа в Конферентната зала на БИЦ на Технически университет – София на открито заседание на научното жури, определено със заповед № ОЖ-5.3-61/04.11.2024 г. на Ректора на ТУ-София в състав:

1. Доц. д-р инж. Юлиян Славейков Велчев – председател
2. Проф. д-р инж. Милена Кирилова Лазарова-Мицева – научен секретар
3. Проф. д-р инж. Александър Богданов Бекярски - пенсионер
4. Доц. д-р инж. Страхил Ангелов Соколов - ВУТП
5. Проф. д-р Кирил Методиев Алексиев – ИИКТ – БАН

Рецензенти:

1. Доц. д-р инж. Юлиян Славейков Велчев
2. Проф. д-р инж. Александър Богданов Бекярски

Резервни членове за журито:

1. Проф. д-р инж. Веска Младенова Георгиева
2. Проф. дтн Тодор Атанасов Стоилов – ИИКТ- БАН

Материалите по защитата са на разположение на интересуващите се в канцеларията на Факултет по Телекомуникации на ТУ-София, блок №1, кабинет № 1254.

Дисертантът е редовен докторант към катедра „Радиокомуникации и видеотехнологии“ на Факултет по Телекомуникации. Изследванията по дисертационната разработка са направени от автора, като някои от резултатите от тях са публикувани.

Автор: маг. инж. Пламен Атанасов Христов
Заглавие: Визуален анализ на поведението на индивиди в киберфизични системи
Тираж: 30 броя
Отпечатано в ИПК на Технически университет – София

I. ОБЩА ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

Актуалност на проблема

Анализът на човешко поведение играе важна роля при интегрирането на машинното обучение и компютърното зрение във физическите среди от реалния свят. Тази интердисциплинарна област е обект на забележителен напредък в последните години, воден от експоненциалния растеж в изчислителната мощност на графичните процесори (GPU), както и в бързото развитие на методите за дълбоко обучение. Чрез оползотворяването на различни подходи от компютърното зрение, компютърните системи могат да откриват, анализират и реагират на човешко поведение във физически пространства.

Един от ключовите компоненти в човешкото поведение е човешката активност, при която анализът се фокусира върху разпознаването и интерпретирането на сложни човешки поведения от видео данни. През последните десетилетия методите еволюират от откриване на прости движения в разбиране на нюансирани действия и дори предсказване на бъдещи такива. Успоредно с това, напредъкът в откриването на човешка поза позволява на системите за наблюдение прецизно моделиране и проследяване на динамиката на човешкото тяло в реално време. Така се изграждат основите за по-задълбочен анализ на човешкото поведение. Въпреки тепърва доказващите се методи за определяне на 3D поза и реконструиране на 3D тяло от 2D изображение, интегрирането на множество гледни точки чрез методите за сливане на ракурси допълнително би подобрило устойчивостта и точността на тези системи. Чрез комбинирането на данни от различни видеокамери или сензори, киберфизичните системи придобиват по-цялостно разбиране на човешкото поведение, което от своя страна отваря много нови възможности в различни приложения от реалния живот. Визуалният анализ на човешка активност продължава да еволюира постепенно и до този момент и в бъдеще можем да очакваме безпроблемно интегриране на компютърната интелигентност в нашия свят, потенциално и драстично променяйки начина по който разбираме и взаимодействаме с околната среда.

В тази дисертация ще се разгледат различни методи и алгоритми за анализ на човешка активност, която е неизменна част от човешкото поведение и е нейн главен представител в научната литература. Методите и алгоритми включват разпознаване на човешка активност, откриване на необичайна човешка активност и откриване на взаимодействие индивид-обект. Тези видове анализ могат да бъдат приложени в киберфизичните системи, откриващи падане на възрастни хора или тези, поддържащи безопасната работа в производствените процеси. Данните, които се получават в тези системи, често са уязвими на припокриване и шум, което изисква въвеждането на допълнителни ракурси. Работата в множество ракурси създава допълнителни задачи за решаване, поради което ще се разгледат два метода за разпознаване на активност в подобна среда.

Цел на дисертационния труд, основни задачи и методи за изследване

Целта на настоящата дисертация е разработване на методи и алгоритми за различни видове анализ на активности с приложна насоченост, както при моноракурсни системи, така и при многоракурсни.

От поставената цел на дисертационния труд произтичат следните основни задачи:

1. Разработване на комбиниран метод и алгоритъм за класификация на човешка активност чрез дълбока невронна мрежа;
2. Разработване на метод и алгоритъм за откриване на взаимодействие индивид-обект чрез дълбока невронна мрежа;
3. Разработване на метод и алгоритъм за откриване на аномална активност чрез дълбока невронна мрежа;
4. Разработване на метод и алгоритъм за разпознаване на човешка активност в многоракурсна система;
5. Функционална проверка на действието на разработените алгоритми, чрез програмна реализация и готови множества данни, за оценка на тяхната ефективност.

Научна новост

В дисертационния труд са предложени методи и алгоритми за анализ на активност в моноракурсни и многоракурсни системи. Предложените методи стъпват на съвременни концепции и подходи в машинното обучение, дълбокото обучение и анализа на времеви последователности. Използвани са съвременни софтуерни инструменти за реализация на алгоритмите, включващи алгоритми за ускорени изчисления и работа с бази данни.

Практическа приложимост

Представените нови методи и алгоритми в този труд са приложими като компоненти от многоракурсни киберфизични системи с човек във веригата. Методът за откриване на необичайна активност може да се приложи в домашни системи за наблюдаване на здравето на индивиди. Методът за откриване на взаимодействие човек-обект може да се приложи в индустриални системи, наблюдаващи и подсигуряващи производствен процес. Методът за сливане на активност могат да послужи за по-качествена оценка на активността в подобни системи.

Публикуване на резултатите от дисертационното изследване

Направените анализи, предложените подходи и получените резултати са представени в общо 4 авторски публикации на престижни международни конференции. Една от публикациите е самостоятелна, а останалите 3 са в съавторство. Налични са общо 14 цитирания, всички от които са в индексирани издания.

Международните конференции са:

1. 2020 XXIX International Scientific Conference Electronics (ET);
2. 2020 28th National Conference with International Participation (TELECOM);
3. 2021 12th National Conference with International Participation (ELECTRONICA);
4. 2021 IEEE 20th International Symposium on Network Computing and Applications (NCA).

Структура и обем на дисертационния труд

Дисертационният труд е в обем от 111 страници формат А4 и съдържа увод, въвеждаща глава, три основни глави, заключение с изложени основни приноси, списък на фигурите, списък на таблиците, списък на използваните съкращения, списък с публикациите по дисертацията и списък на използваната литература. Изложението на дисертационния труд, направено в 3 глави, съдържа 31 фигури и 9 таблици. Използвани са 103 литературни източници.

Номерата на фигурите, таблиците и математическите изрази в автореферата съответстват на тези в дисертационния труд.

II. СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. Анализ на състоянието на проблема по литературни данни

1.1 Визуален анализ на поведение, общи постановки на проблема и насоки за решаването му

Методите за визуален анализ на поведението са еволюирали значително с развитието на технологиите. Традиционните подходи са разчитали главно на човешкото зрение и ръчното анотиране на човешкото поведение. С навлизането на компютърното зрение и машинното обучение, тази сфера се усъвършенства, позволявайки автоматизиран анализ на голям обем видеоданни. Главните визуални модалности, използвани за анализ на поведение включват стандартните цветови видеоданни, дълбочинни и инфрачервени видеоданни, извлечени от дълбочинни сензори, които подобряват 3D разбирането на заснетите сцени.

1.2 Представяне на активности и видове представяне

В литературните източници действие (action) и активност (activity) са най-често взаимозаменяеми термини и означават едно и също. Значението на самите думи подсказват, че активността е състояние на заетост, в която са включени поредица от действия за период от време. Задачите на анализ и при двете значения представляват най-често класификация, където всяка единица е последователност и с други сходни по дължина

и сложност последователности съставляват дадено множество данни. Известните множества данни [1,2,3] обхващат голям диапазон от различни активности или действия като те самите могат да бъдат нещо просто от типа "козируване" или нещо по-сложно като "скачане" и "бягане", където има движение на тялото тяло.

В определени научни публикации авторите дефинират активностите като последователност от множество действия във времето, което им помага за по-нататъшно разпознаване и класификация на същите активности. Такъв пример е научният труд на Liu et al. [4]. В него примери за активност включват "ранна сутрин", "чистене" и "време за кафе". Примери за действия при тяхното множество данни са: "ходене", "ставане", "сядане", "досягане", "отваряне (на врата)".

Голямата част от множествата данни от активности, използвани в литературните източници, се състоят от отделени откъси на кратки действия. Тези откъси са отделени или анотирани ръчно от авторите без да се вземе предвид контекста, тъй като задачите, които решават, са в офлайн обстановка [5]. Някои множества се състоят от дълги последователности с много активности, с анотирани начало и край, които имат за цел предварително откриване на активностите преди по-нататъшен анализ. И при двата вида множества данни активностите са еднородни, но разликата е в това дали са част от по-голяма последователност.

За целите на дисертацията ще се дефинират три нива на активност:

- Активност - състояние на заетост, включващо поредица от прости активности, продължаваща в дълъг период от време. Във всяка активност има голяма междукласова дисперсия между класовете на действията;
- Действие - последователност от множество движения на тялото (общи и елементарни). Във всяко действие има малка вътрешнокласова дисперсия в междукадровите разлики. Нарича се още проста активност;
- Движение - съвкупност от елементарни отмествания на тялото. Те се групират в клъстери или класове на базата на определени компоненти на тялото.

Оттук нататък ще се използва терминът "активност" за общо наименование на "активност" и "действие".

Представянето на човешки активности е решаваща част в анализа на активностите, тъй като тя директно афектира производителността в последващото им разпознаване и разбиране. През годините изследователите са разработили разнообразни подходи за представянето на човешки активности, всеки със своите ползи и недостатъци. Литературните източници описват три главни категории на представяне - холистично, представяне чрез локални признаци и траектории, и чрез скелетонна поза.

Подходите при холистичното представяне взимат предвид човешкото тяло или пълната сцена като единствено и цялостно образувание, което може да се анализира. Тези методи обикновено извличат глобални признаци от цяло изображение или видео, анализирайки общата външност и тенденциите на движение, асоциирани с различните движения.

За разлика от холистичните методи, локалните признаково-базирани подходи представят активностите като колекция от пространствено и времево локализирани признаци или траектории. Тези методи обикновено са по-устойчиви на частично припокриване и резки промени във фона. Основополагаща разработка в тази категория са пространствено-времевите интересни точки (STIPs). STIPs са локални признаци - точки, открити в области от видеото. Там стойностите на изображението имат изпъкващи локални вариации в пространството и времето и движението е променливо.

Представянето на активност чрез скелетони привлича внимание през последните години поради способността да отделя структурната информация на човешките тела и нейната невариантност при вариации в сцената. Този вид подходи представя човешките активности като поредица от позиции или стави на тялото, в двумерно или тримерно пространство.

Появяването на широко достъпни дълбочинни сензори като Microsoft Kinect на пазара революционизира оценката на 3D поза. Дълбокото обучение значително допринася за напредъка в сферата на определяне на позата, както в 2D, така и в 3D.

1.3 Обзор на методите и алгоритми за анализ на активности

Анализът на активност е общата област, включваща задачи от всякакво естество, свързани с човешката активност. Съществуват различни видове анализ на активност. Спрямо взаимодействието в активността това може да е анализ на самостоятелна активност, активност между двама индивиди и групова активност (много индивиди). Спрямо вида машинно обучение това може да е задача, свързана с обучение чрез учител или без такъв, която може да се приложи при повечето други времеви последователности. Във Фиг. 1.2 е направена пълна категоризация на видовете.

Фигура 1.2: Видове анализ на активности

Да приемем, че даден индивид извършва предварително определен набор от M на брой активности A, които могат да бъдат обозначени като:

$$A = A_0, A_1, ..., A_{M-1} \tag{1}$$

Една камера или сензор се използват за извличане на списък от атрибути S от R на брой времеви последователности в рамките на интервал от време $I = [t_\alpha, t_\omega]$ за разпознаване на тези активности:

$$S_{atr} = S_0, S_1, ..., S_{R-1} \tag{2}$$

Целта при разпознаването на човешка активност в най-общия си вид е да се открие подходящо времево разделяне $[I_0, ..., I_{R-1}]$ от интервали I, като се използва списъкът с атрибути S_{atr} и колекция от класове R, описващи активностите, които са били извършени в рамките на всеки интервал I_j. Това предположение показва, че разпределението на времето I_j е непрекъснато, не се припокрива и $\bigcup_{j=0}^{R-1} I_j = I$.

Разпознаването човешка активност (HAR) се среща много често в научните изследвания, използващи дълбоко обучение. Това се дължи на многобройните обемни множества данни (dataset), съпроводени от състезателни авторски еталонни конфигурации и тестове (benchmark), които допринасят за прогреса в невронните методи и архитектури. Навлизането на сензорите Kinect на пазара дава тласък на научния анализ на активности от скелетони поради лесното им записване.

Въпреки богатия набор от статии, свързани с общото разпознаване на активности, съществуват по-приложни проблеми, които се нуждаят от специализирани методи. Такъв е проблемът с откриването на необичайна човешка активност.

Откриването на необичайна човешка активност е съществена област на изследване със своите приложения в здравеопазването, сигурността и подпомогнатия живот. Тази област обхваща различни подобласти като откриване на падане, анализ на походка и откриване на обща аномалия. Откриването на падане, специално, е интересна тема в изследователските и индустриалните среди, набираща своята популярност от 2019-та година, особено в развитите страни с висок стандарт на живота, но които са придружени, за съжаление, от проблема със застаряващото население [42].

Когато индивидите се опитват да разберат околната среда, те го правят чрез наблюдаване на това как други хора взаимодействат с обекти. Целта на откриването на взаимодействие индивид-обект (Human-object interaction, HOI) е да определи правилно индивидите, обектите и активностите, които се случат между тях, в статичен момент от времето или последователност във времето. HOI е ключова сфера на изследване с приложения в роботиката, интерфейсите човек-компютър и киберфизичните системи.

1.4 Обзор на видеосистемите за проследяване и анализ на активностите

За бърза реакция при различни наблюдавани събития, за тях е нужно физическо присъствие през 24 часа от денонощието. Затова са нужни автоматизирани видеосистеми, които улесняват процеса на наблюдение. Главната роля на автоматичните системи

е, да подпомогне оператора да извърши разнообразните си задачи, свързани с наблюдението. Те могат да бъдат свързани с различни приложения като например контрол на трафика, предсказване на инциденти, предотвратяване на престъпления, засичане на аномалии и обществена сигурност. Анализът на човешка активност и поведение намира приложение при системите в затворени пространства - това могат да бъдат системите, откриващи падане при грижата за възрастни или наблюдаващи поведението на пациенти.

Всяка видеосистема, включваща едно видеоустройство, се нарича моноракурсна. В литературните източници за този общ тип системи няма обособена категория, но съществуват отделно разработки, които представят задачите, които тези системи решават, както и тези, представящи мрежовата част и базата данни.

При класическите методи за анализ на активности във видеопоследователности е честа практика отделянето и проследяването на индивидите с прозорци при записи на далечна сцена, където могат да бъдат включени и други индивиди. Този подход е сходен с "top-down" подходите при изчисляване на поза, където първо се открива и отделя с прозорец субекта.

При сензорите Kinect и дълбоките невронни мрежи за определяне на поза тази предпоставка е гарантирана в някакви граници, а именно - броят проследявани индивиди и резолюцията на изображението. При излизането извън тези граници производителността драстично спада или процесът прекъсва тотално.

Всяка видеосистема, включваща две или повече зрителни устройства, се нарича многоракурсна. Това може да са обикновени RGB камери или дълбочинни сензори. Многоракурсните системи са подходящи за задачи в открити пространства като анализ на спорт и изчерпателно наблюдение на големи площи поради справянето им с припокриванията и проследяване на обекти от много зрителни ъгли.

Многоракурсните системи се разделят на такива със споделено зрително поле (с припокриване на зрителните полета на видеокамерите) и такива без припокриване. В зависимост от заснетото пространство, те могат да бъдат в открито и закрито пространство.

Главното предимство на многоракурсните системи пред моноракурсните се изразява в допълнителното количество данни, от различни гледни точки, което те предоставят. Чрез тези данни може да се реши проблемът с припокриването и проследяването на индивидите. Общите проблеми, които се пораждат при анализа на активности в една многоракурсна система са следните:

- Работа с много ракурси - избор на един ракурс или обединяване на няколко ракурса;
- Работа в реално време - избор на методи и алгоритми, работещи в реално време;
- Работа с много индивиди - отделяне на индивиди и обекти в контекста на един кадър;
- Работа с дълги последователности - сегментиране на активности или действия във времето.

1.5 Данни при анализ на човешка активност

Събирането на данни, независимо дали се отнася до придобиването на сензорни сигнали или видеоклипове, е важна и основна част на всяка система, анализираща човешка активност. Данните могат да бъдат извлечени от носими сензори и видеосензори.

1.5.1 Данни, извлечени от носими сензори

Еволюцията на Интернет на нещата (Internet-of-things, IoT) и мобилните компютри през последните години създава перфектна среда за използването на различни типове сензори (носими и видео). Носимите сензори са най-разпространените сензори при анализа на човешка активност. Те могат автоматично да открият много различни активности като "сядане", "бягане" и "спане". Трите стандартни носими сензора са акселерометър, магнитометър и жироскоп. Освен да се носят като отделни устройства, те могат да се интегрират в преносими устройства, като смартфони, смарт часовници, смарт ленти, очила или каски. След това могат да бъдат открити човешки активности чрез измерване на разликите в сигнала преди и след активност.

Друг вид сензори са сензорите за обекти. Това са сензори, прикрепени към конкретен обект, с цел идентифициране на активности, свързани с този обект. Докато носимите сензори измерват директно човешката активност, сензорите за обекти откриват движението на конкретни обекти, за да направят извод за човешката активност. Сензорите за обекти се използват по-рядко от сензорите за носене поради високите разходи и предизвикателствата при настройката.

За разлика от носимите сензори и сензорите за обекти, сензорите за околната среда обикновено се поставят в средата на индивида, за да измерят точни данни, свързани с основните параметри на околната среда като влажност, температура, CO2 и прахови частици. Екологичните сензори се използват за наблюдение на промените в околната среда при появата на физическо движение. Тъй като сензорите за околната среда са силно чувствителни към промените в околната среда, приемането на подходящи сензори за околната среда трябва да бъде внимателно планирано въз основа на активностите [48].

Данните, получени от носимите сензори, се представят като последователности от различни по дължина вектори.

1.5.2 Зрително-базирани данни

Зрително-базирани данни (или видеоданни) са ключови при анализа на активност в системите за видеонаблюдение. В контекста на анализа на човешка активност, те могат да бъдат разделени на RGB данни и RGB-D данни.

RGB данните съдържат червени, зелени и сини честотни ленти във видимия спектър, които могат да бъдат записани с помощта на камери, оборудвани с обикновен CMOS сензор (Complementary metal-oxide semiconductor). Това прави тези данни широко достъпни и с характеризиращи се с богати текстурни описания на проследяваните индивиди. Въпреки това, CMOS сензорът има ограничен обхват, зависим е от калибрирането и до голяма степен от условията на околната среда, като осветлението и припокриването във фона.

Благодарение на разработването на сензори за дълбочина, анализът на човешка активност може да се извърши по-точно. Освен основните RGB данни, RGB-D камерите заснемат и информация за дълбочината в пространството, която може да помогне на алгоритмите да разпознават по-точно човешките движения. Друго голямо предимство е приложимостта на алгоритмите за изчисляване на скелетон от човешкото тяло, което решава проблема с откриването и извличането на признаци от проследяваните индивиди. Поради това, RGB кадрите, постигат по-ниска точност при разпознаване на активност в сравнение с RGB-D кадрите, тъй като мултимодалността на RGB-D данните предоставя допълнителна информация под формата на дълбочинен канал.

Основно предимство на RGB-D сензорите и данните, получени от тях, е устойчивостта при промени в осветеността, цвета и текстурата, както и способността им да функционират в тъмна среда. Дълбочинните данни обаче имат ниска разделителна способност и могат лесно да бъдат повлияни от някои материали, които абсорбират светлината. Друг недостатък е изчислителната сложност, която нараства при анализ на големи множества данни [48].

Видеоданните, като цяло, са по-гъвкави по отношение на анализ и приложения спрямо данните от носими сензори, поради по-голямата размерност и информация, която може да се извлече от тях (например скелетони или позициите на обектите в сцената). Събирането на видеоданните спрямо това на сензорните данни е неинвазивно, т.е. не влияе върху действията на индивидите. В същото време, тъй като камерите имат ограничен зрителен ракурс и зрително поле, и са стационарни, те не могат да проследяват обекти извън заснемената сцена. Независимо от това, този проблем се решава при въвеждането на повече камери, или създаването на многоракурсна система.

1.5.3 Сензори Kinect

Kinect е комбиниран видеосензор на движение, произвеждан от Microsoft и първоначално се появява на пазара през 2010. Неговата технология включва RGB камери, инфрачервени прожектори и детектори на дълбочина, които са осъществени или чрез изчисляване на структурирана светлина(structured light) или чрез времето на полет на лъчите (time-of-flight). Също така включват микрофонна решетка. Всичко това, заедно със софтуера, предоставен от Microsoft, позволява извършване на разпознаване на жестове и говор, както и изчисляването скелетони на индивиди в реално време.

Второто поколение сензори Kinect (2013 – 2014) работят чрез Time of Flight(ToF) и използват подхода на интензитетна модулация на непрекъсната вълна, който е най-често използван при ToF устройства. Неговата идея е да се осветява активно сцената, която се наблюдава, с интензитетно-модулирана, периодична светлина от инфрачервения спектър. Поради разстоянието между камерата и обекта, както и ограничената скорост на светлината c, се предизвиква дадено времево отместване $\phi[s]$ в оптичния сигнал, което е еквивалентно на фазово отместване в периодичния сигнал. Това отместване се открива във всеки сензорен пиксел. Времевото отместване може лесно да се трансформира в разстояние сензор-обект, тъй като светлината трябва да измине разстоянието два пъти.

Всяко Kinect устройство идва със свой комплект за разработка, позволяващ проектиране на приложения, които използват данните, получени от сензора. Тези комплекти имат интуитивен интерфейс, който улеснява извличането на следните данни:

- Цветови кадри - Предоставят се от RGB камерата на сензора и могат да бъдат достъпени от множество цветни формати;

- Инфрачервени кадри - Представляват инфрачервени изображения, заснети от дълбочинната камера на сензора;

- Дълбочинни кадри - Предоставят дълбочинна информация от дълбочинната камера на сензора. Дълбочината се дава под формата на милиметри разстояние от равнината на камерата до най-близкия обект в определена пикселова координата;

- Скелетонни кадри - Те разкриват цялата информация относно човешките индивиди, които са в зрителния обхват на сензора. Предоставят координати на стави и техните ориентации на максимум 6 индивида при Kinect 1 и 2 и без ограничение в броя при Azure Kinect (при понижаване на производителността и кадровата скорост). Един скелетон се състои от 20/25/32 стави(Kinect V1/V2/Azure). Всяка става носи информация за X, Y, Z координати в 3D сензорното пространство, както и нейната ориентация.

Комплектите за разработка работят с 3 координатни системи – цветова, дълбочинна и сензорна. Цветовата и дълбочинната имат за координатно начало (x = 0, y = 0) горният ляв ъгъл от съответните им изображения. Долният десен ъгъл съответно зависи от резолюцията на камерите им. Сензорните координати се използват от инфрачервения сензор на Kinect, за да открият 3D точките на ставите в пространството. Това са координатите, които се използват за позициониране на ставите или точките на лицето. Сензорното пространство се отнася до 3D координатната система в Kinect.

За да се изобрази скелетон е нужно да се проектират ставите му в 2D пространство, т.е. върху цветово или дълбочинно изображение. Комплектите за разработка предоставят готови функции за тази цел.

1.7 Дефиниране на целта и основните задачи на настоящата дисертация

Целта на настоящата дисертация е разработване на методи и алгоритми за различни видове анализ на активност с приложна насоченост, както при моноракурсни системи, така и при многоракурсни. Ще се използват свободнодостъпни множества данни, съдържащи анотирани и сегментирани активности и действия(прости активности) от скелетони, които са извлечени от видеосензори Kinect.

От поставената цел на дисертационния труд произтичат следните основни задачи:

1. Разработване на комбиниран метод и алгоритъм за класификация на човешка активност чрез дълбока невронна мрежа;

2. Разработване на метод и алгоритъм за откриване на взаимодействие индивид-обект чрез дълбока невронна мрежа;

3. Разработване на метод и алгоритъм за откриване на аномална активност чрез дълбока невронна мрежа;

4. Разработване на метод и алгоритъм за разпознаване на човешка активност в многоракурсна система;

5. Функционална проверка на действието на разработените алгоритми, чрез програмна реализация и готови множества данни, за оценка на тяхната ефективност.

1.8 Описание на работната среда

Предложените методи и алгоритми в следващите глави предполагат като предпоставка затворена система от няколко сензора Kinect с припокриващи се зрителни полета и разположени на равни ъгли спрямо центъра на заснетата сцена. Подобна система е предложена в научната публикация [D1].

В следващите глави ще бъдат описани задачи, **включващи един индивид** и където активностите са **предварително сегментирани във времето**. Проблемът с вътрешното и външното калибриране на сензорите ще се предположи, че е решен. Различните множества данни, върху които ще се извършват експериментите, са заснети от системи като [D1].

ГЛАВА 2. Методи и алгоритми за анализ на активности от скелетонни данни чрез дълбоки невронни мрежи

2.1 Комбиниран каскаден метод за разпознаване на активност чрез CNN и SVM

Методът се състои от пет последователни стъпки: нормализиране на последователностите от скелетони, изравняване на дължината им, представянето им по два различни начина, обучение на CNN с тези представяния и използване на резултатите от предпоследния слой на невронната мрежа за обучение на машина за опорни вектори (SVM). На Фиг. 2.4 е показан алгоритъмът към представения метод.

Дефинират се два варианта на вход на дълбоката невронна мрежа: тензор от нормализираните скелетон-вектори и вектор от разлики между скелетон-векторите. Тензорната форма се реализира чрез формиране на тензор на трите матрици с размери $N \times K$ (N - брой кадри, K - брой стави) и стойности на компонентите, съответващи на всяка координата x, y и z. При вектора от разлики всеки компонент е Евклидово разстояние между два скелетон-кадъра. Скелетон кадрите са представени като вектори, в които компонентите (K на брой) са нормализираните Евклидови разстояния между всяка става и центърът на тежестта G на съответния скелетон.

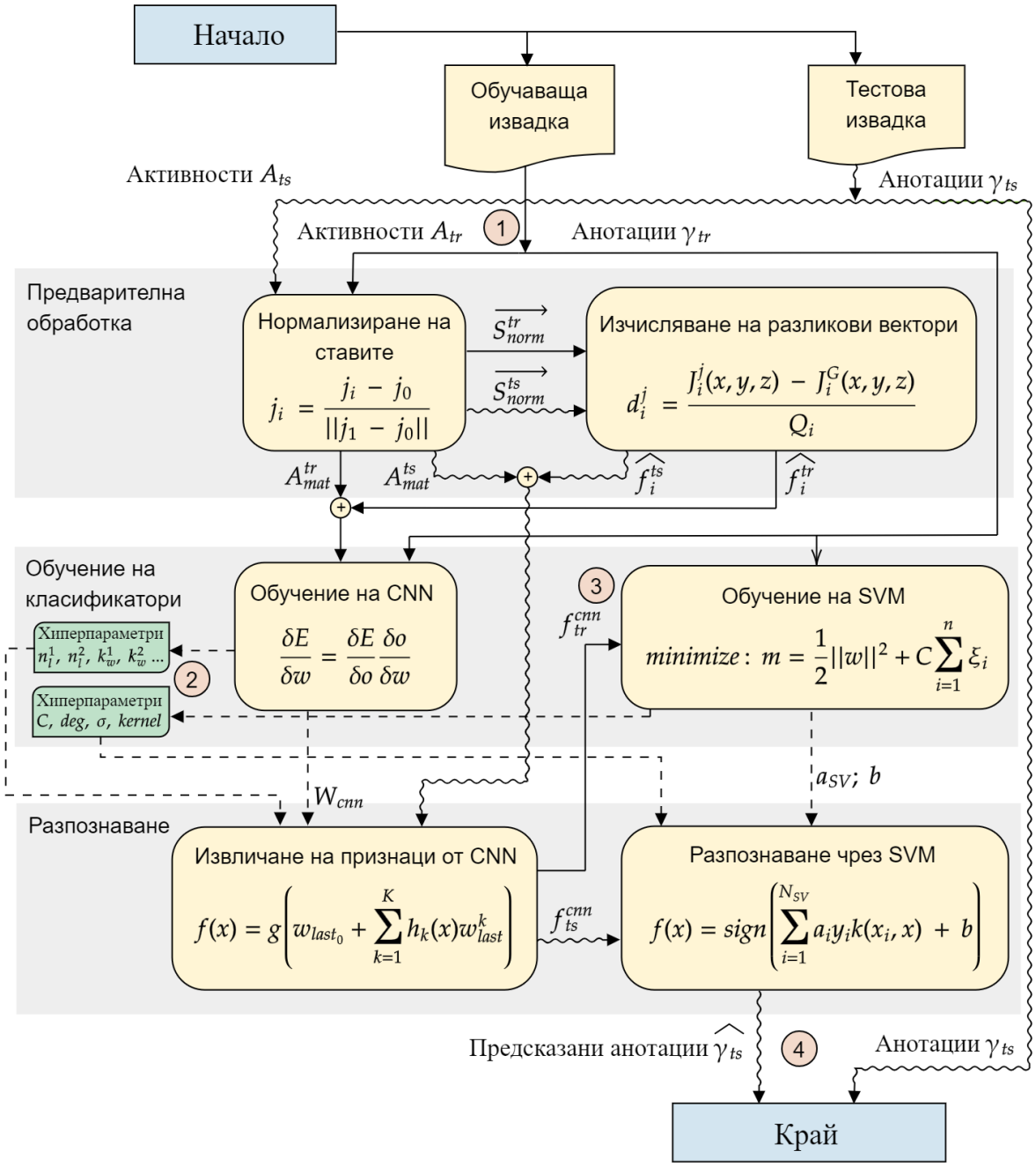
Дадена става i на скелетон се представя като 3D вектор $\overline{j_i}(x, y, z)$, за кадър от активността A_m, състояща се от N на брой кадри. Центърът на ставите J_G^j за този скелетон се изчислява като средна стойност от 3D позициите на K-те на брой стави в скелетона. За да станат разликовите вектори, с разлики между ставния център и всяка става, мащабно инвариантни, се въвежда нормалиращ фактор Q, представящ тяхната средна стойност. Всяка разлика d_i^j се разделя на нормализиращия фактор. Векторът на позата $\widehat{f_n^j}$, асоцииран с n-тия кадър, съдържа K на брой разлики. Разликовият вектор

се формира чрез изчисляване и наслагване (конкатениране) на Евклидовите разстояния между всеки два вектора на поза, достигайки размер $\frac{(N*(N-1))}{2}$.

Оптималните параметри се определят чрез Бейсова оптимизация и са описани по-долу:

- Брой на филтрите на двата конволюционни слоя на горния блок на CNN;
- Размер на ядрото на филтрите на двата конволюционни слоя на горния блок на CNN;
- Брой неврони на единичния изцяло-свързан слой на долния блок;
- Брой неврони на последния изцяло-свързан слой;
- Вид ядро на SVM-ът – (линейно / полиномно / гаусово);
- Регуляризиращ параметър C на SVM;
- Ред на полинома на полиномното ядро на SVM;
- Параметър гама за гаусовото ядро на SVM.

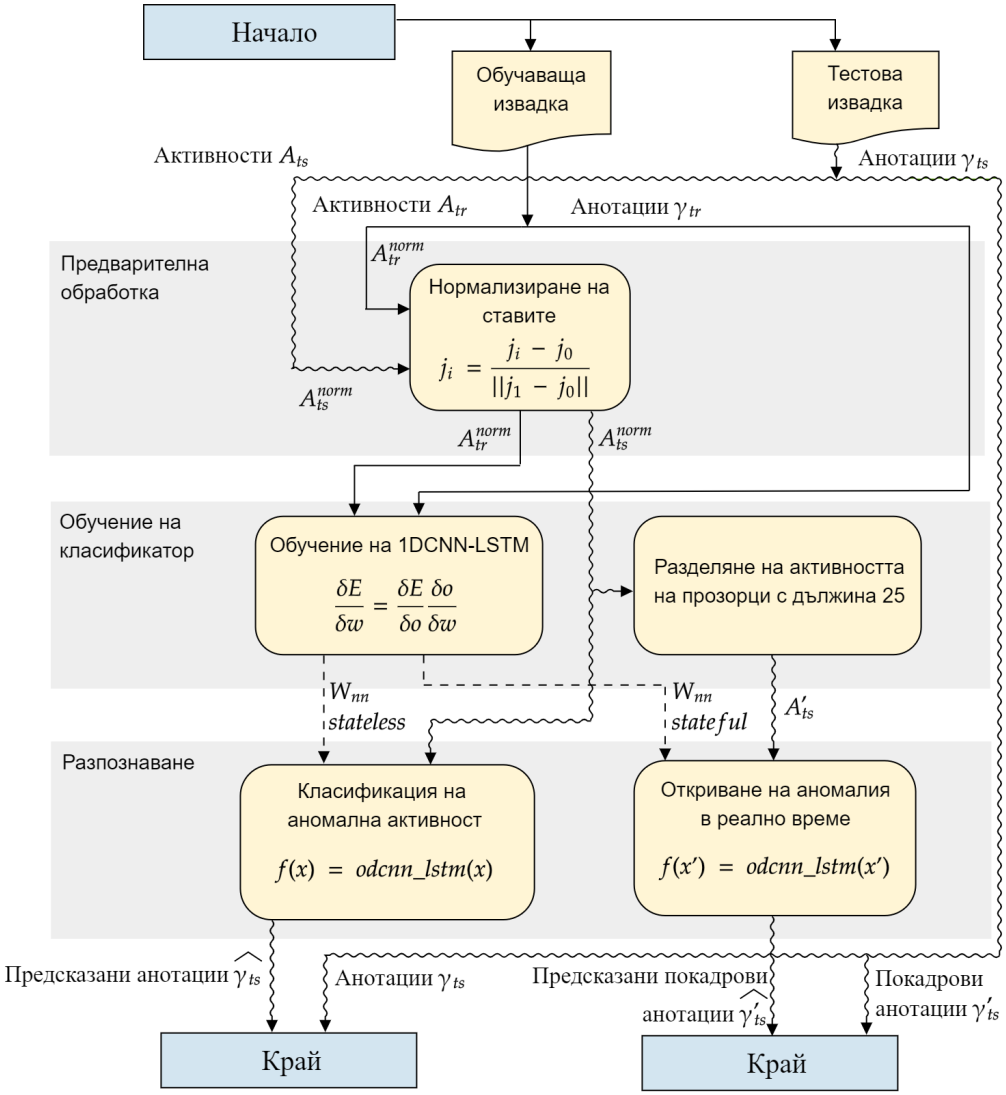
След като CNN-ът се обучи върху всички данни, те се подават още веднъж на модела, за да се извлече изходът от предпоследния слой под формата на вектори с размер 384, с които се обучава SVM-ът. Векторите са нормализирани в диапапазона $[0, 1]$, тъй като са изходни резултати от сигмоидалния слой на невронната мрежа.



Фигура 2.4: Блок схема на предложения алгоритъм за разпознаване на активност чрез CNN-SVM. Включени са режим на обучение и режим на тестване на работата. Плътните линии са за обучение, вълнообразните са за тестване. С цифри в кръг е обозначена поредицата на операциите. Прекъснатите линии символизират преход на параметри или хиперпараметри.

2.2 Метод за откриване на необичайна човешка активност в реално време чрез 1DCNN-LSTM

Методът представлява модел на дълбоко обучение, който съчетава 1D конволюционна невронна мрежа (1D CNN) и мрежа с дългосрочна памет (LSTM). Данните са набор от последователности, всяка от които се състои от признаковите вектори, компонентите на които се отнасят до 3D координатите на ставите на скелета. Във Фиг. 2.9 е представена блок-схема на алгоритъма за откриване на необичайна човешка активност.



Фигура 2.9: Блок-схема на алгоритъма за откриване на аномална човешка активност. Включени са режим на обучение и режим на тестване на работата. Плътните линии са за обучение, вълнообразните са за тестване.

Мрежата, представена тук, се състои от 3 последователни 1D CNN слоя. Всеки слой е последван от максимално-обединяващ слой (Max Pooling) и Spatial Dropout. Конволюционните слоеве имат ядра на филтъра с ширина 3, 5 и 7 кадъра и каузална,

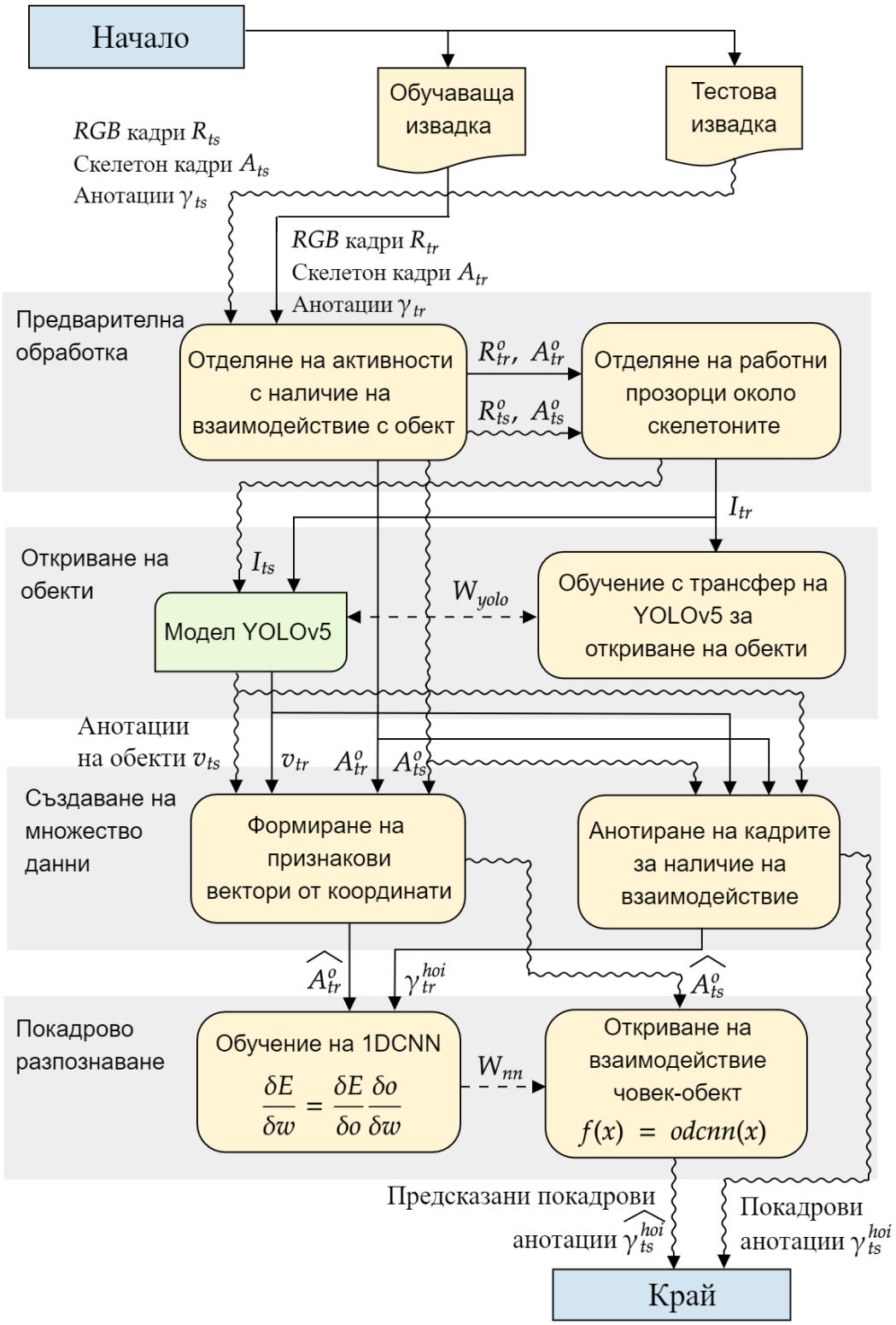
разширена структура. Дължината им остава същата като дължината на вектора (75 при 25 стави). Каузалните конволюции взимат предвид само елементи в текущи и отминали моменти от времето, тоест текущият изходен елемент от операцията конволюция попада на последната (във времето) позиция от ядрото на филтъра. Това гарантира, че няма да се наруши преходът на информация в реално време. Разширените конволюции изчисляват претеглена сума само на елементи през даден брой позиции в ядрото, което предотвратява усложняването на модела при по-големи размери на ядрото. В предложения метод конволюционният слой с размер на ядрото 7 има степен на разширяване 3, което означава, че се вземат елементите през три позиции (първи, четвърти и седми елемент). Spatial Dropout е вид операция за регуляризация, която премахва цели признакови карти по времевата димензия, за да избегне образуване на корелация между съседни карти. Дялът на премахнатите карти е избран да бъде 0.1.

Резултатът от 1DCNN се предава на LSTM, който връща признаков вектор след последния кадър. След това този вектор се предава на изцяло-свързан слой, който изчислява вероятност за аномалия между 0 и 1 за цялата последователност. По време на прогнозиране в реално време, LSTM се превключва в режим на запазване на състоянието (stateful), така че да получава и обработва данните кадър по кадър. След това към модела се подава последователност от скелетони, извлечени от плъзгащ се работен участък, преминаващ през пълната последователност и воден от текущия кадър. Размерът на работния участък зависи от конволюционните и максимално-обединяващите слоеве. За тази реализация той е избран да бъде 25, така че след обработка да бъде сведен до един вектор от тези слоеве, който ще бъде предаден на LSTM. По този начин за всеки скелетон-кадър 1DCNN връща като резултат извлечените признаци за натрупаната последователност до момента.

2.3 Метод за откриване на взаимодействие индивид-обект чрез 1D конволюционна невронна мрежа

Алгоритъмът започва с отделяне на активностите-екземпляри от дадено множество данни, където има взаимодействие на индивид с обект. Тези екземпляри включват RGB видеопоследователности и последователности от скелетон-кадри. Ставите на скелетоните са в 2D формат, тъй като са проектирани върху RGB кадрите. Видеоклиповете, записани с Kinect v2, са с резолюция 1920×1080 пиксела и честота на опресняване $30Hz$. Местоположението на обектите се открива или проследява по време на видеоклиповете чрез невронния модел YOLOv5 и алгоритъма CSR-DCF. YOLOv5 предварително се обучава с отделени кадри от видеоклиповете, където обектите са ръчно анотирани. В комбинация с информацията за позата на тялото, предоставена от Kinect v2, се формират последователности от вектори за всяка активност, които представят движението на ръцете на индивида спрямо позицията на обекта във времето. За задачата прогнозиране на наличие на взаимодействие за всеки кадър от времето се използва 1DCNN.

Във Фиг. 2.14 е представена блок-схема на алгоритъма.



Фигура 2.14: Блок-схема на представеният алгоритъм за откриване на взаимодействие индивид-обект. Включени са режим на обучение и режим на тестване на работата. Плътните линии са за обучение, вълнообразните са за тестване.

Дефинират се следните променливи, свързани с дефиницията за взаимодействане:

- Минимално изминато разстояние d_{start} на даден обект за започване на взаимодействие в даден интервал от време;
- Максимално изминато разстояние d_{end} на даден обект за приключване на взаимодействие в даден интервал от време.
- Максимален интервал от време n_{start} (в брой кадри), в който обектът може да измине разстояние d_{start}, започвайки взаимодействие;
- Максимален интервал от време n_{end}, в който обектът спира да се движи и не изминава разстояние повече от d_{end}, приключвайки взаимодействие.
- Максимално разстояние d_{match} за съпоставяне на един и същ обект между 2 кадъра;
- Максимален интервал от време n_{loss}, в който обектът има разлика в координатите от ставата на ръката и проследените от YOLOv5/CSR-DCF, при която имаме край на проследяването на дадения обект и за кадрите до този момент от момента на надвишаване на d_{match} казваме, че няма взаимодействие.

Променливите за разстояние d_{start}, d_{end}, d_{match} се измерват в 2D Евклидово разстояние, а променливите за интервал от време n_{start}, n_{end}, n_{loss} - в брой кадри.

Прави се обучение с трансфер с данни - всеки 15-ти кадър от видеозаписите. Тъй като YOLOv5x е обучен на изображения с разделителна способност 640×640 пиксела, от извадката кадри се изрязва работен участък с такъв размер, като същевременно се побира вътре скелетона на индивида в сцената.

За всеки RGB кадър се извлича работен прозорец с размери 640×640 пиксела по начин, по който скелетонът на проследявания индивид е разположен в него, без да се вземат предвид ставите на долната част на тялото от коляното и надолу, тъй като в предложения сценарий се използват само ръце. Това позволява скелетоните на индивидите да се представят в непълен размер, тъй като представлява интерес само областта от изображението около взаимодействието на обекта и индивида.

При неуспешно откриване на обект от YOLOv5, последната му известна позиция може да се използва за проследяването му, следователно е въведен CSR-DCF тракер, за да запълни празнините между измерванията.

След намиране на основните ground truth данни, стъпката на обучение може да започне. За тази цел се използва 1DCNN с подхода на плъзгащия се работен прозорец.

Тъй като за дадена последователност е необходим само локалният контекст, той е зададен да бъде 30 кадъра. По този начин 1DCNN ще изчисли етикет само за един кадър, а не за цялата последователност. Всяка времева стъпка е вектор, който съдържа 3D координатите на ставите на ръцете и китките на скелетона, както и 2D работния прозорец на обект. Проблемът с множеството обекти, както и индивидите в сцената, се решава чрез създаване на отделен поток за всяка двойка обект-индивид. Всяка последователност се разделя на прозорци от 30 кадъра и се предава към мрежата. Резултатът е точно един вектор, който след това се свързва с изцяло-свързан слой с един неврон. Този неврон генерира оценка за взаимодействие с активираща функция на сигмоида.

ГЛАВА 3. Методи и алгоритми за анализ на активности от скелетонни данни в многоракурсни системи

3.2 Метод за разпознаване на многоракурсна човешка активност чрез оптимална поза

За да се преобразуват скелетон-векторите в мащабно и пространствено инвариантни, те трябва да се нормализират. За този метод те отново се транслират в обща координатна система с център - ставата на кръста, но този път ставите се представят като единични вектори (unit vector). След като се транслира всяка става спрямо общото начало на координатната система, костите на скелетона се представят като вектори, всеки от който е представен от става-родител и става-наследник. След това те се преобразуват в единични вектори чрез изваждане на координатите на родителните стави от дъщерните и разделяне на получения вектор на неговата дължина(норма).

Най-накрая скелетонът се реконструира с костите, представени от единични вектори като итеративно се събират координатите на ставите, започвайки от ставата на кръста и свършвайки със ставите на главата и крайниците. Ъглите на ставите спрямо родителните им стави се запазват, докато костите вече имат дължина 1. За по-добра представителност, единичните вектори се умножават с произволни, но реалистични дължини на костите (единствено при визуализация).

Поради десинхронизацията в успоредно заснемащи дълбочинни сензори, дължаща се на закъснение в изчислението, както и субективност при етикирането на сегменти активност, последователностите от активности трябва да бъдат синхронизирани покадрово. За тази цел тук се използва алгоритъма Dynamic Time Warping (DTW).

Чрез използване на най-малката поредица измежду различните ракурси като основа, скелетоните от другите поредици, които са съпоставени само с единствен скелетон от основната поредица, се пренебрегват, като се остава само първият срещнат скелетон.

Позата на един скелетон се дефинира като посоката, към която той е обърнат спрямо заснемащия сензор, измерена в Ойлерови ъгли (yaw, pitch, roll).

Оттук, оптималната поза се дефинира като поза с координатна система, която е най-сходна с координатната система на дълбочинния сензор, но със z-оста насочена спрямо сензора (с обратен знак). Тази дефиниция се базира на предположението, че скелетонът, който гледа в посока към сензора, ще има най-малкото припокривания и съответно най-малко шум при изчисляването му. Избирането на един скелетон на всеки времеви кадър от множество ракурса също редуцира множеството данни V на брой пъти при такова с V на брой ракурса.

За да се оцени метода преди и след сливане се използват скрити модели на Марков (HMM) за класификация. Класификаторът представлява правило за решение на базата оценките от няколко HMM.

3.3 Метод за разпознаване на многоракурсна човешка активност чрез филтър на Калман

Покадровото сливане на активност чрез оптимална поза е бърз, но неизгладен метод за сливане на позите. Главният му недостатък е приемането на цялата поза като

истинна, без да се вземе предвид шума от самите стави. Освен това, тъй като имаме индивидуално избиране на поза във всеки кадър, не се вземат предвид предходните пози и това би направило почти всяка възможна човешка поза еднакво вероятна ако не знаем пълния контекст. В този ред на мисли в този раздел ще бъде разгледан метод за сливане чрез филтър на Калман, който е класически метод за филтриране и сливане на сензорни данни.

Филтърът на Калман е алгоритъм за предсказването на променливо състояние във времето като използва априорните измервания за тези променливи на всяка времева стъпка. В допълнение взима предвид и шумът на тези измервания.

Сливането на данните е базирано на оптималната поза, представена в 3.2. Главният проблем на метода с оптимална поза е, че създава "колебливост"в цялата активност, дължаща се на рязко сменящи се избрани ракурси, и изразяваща се като шум при по-нататъшното ѝ филтриране. За да се намали до минимум този ефект списъкът с етикети на ракурси V_{map} се изглажда чрез осредняващ филтър с работен участък от 5 елемента, където последния елемент от участъка се променя (каузален филтър).

Скелетонните кадри успоредно се избират спрямо списъка с етикети и се филтрират с филтъра на Калман, с избрани дисперсия на измервателния шум R_{var} 0.05 и на процесовия шум Q_{var} 100, които са измерени в [90]. Измервателният шум в скелетоните включва грешките в изчислението им, което може да се дължи на припокривания на части на тялото, както и шум от изчислената дълбочинна карта. По време на процеса на сливане, данните от другите ракурси също се взимат предвид като участват в стъпката на обновяване при филтъра на Калман, където допълнително променят вектора на състоянието и ковариационната му матрица. Това последователно обновяване на състоянието е класически метод за сливане на данни от сензори и е описано в [91]. Тъй като скелетоните от съседни ракурси са с различна ориентация, те биват ротирани чрез алгоритъма на Кабш [92] за намирането на оптимална матрица на ротация между две поредици от 3D точки в обща координатна система. 3D точките, на базата на които се изчислява матрицата, са тези на торса, раменете и бедрата, тъй като те са най-устойчиви на шум.

Обикновено измервателният шум се измерва при сравнението на филтрирани данни с ground truth данните, като ако данните са скелетонни, за ground truth служат moCap данни с заснети от прикрепени маркери, където грешката е незначителна. За множествата данни, които се използват в тази глава, не съществуват ground truth данни. При предположението, че оптималните скелетони имат най-малко шум, се измерва релативен шум на ротираните неоптимални скелетони спрямо оптималните под формата на дисперция R'_{var}. Тази второстепенна дисперсия се събира с основната R_{var} когато неоптималните скелетони участват в стъпката на обновяването.

ГЛАВА 4. Експериментални резултати от разработените алгоритми

4.1 Експериментални изследвания на методите към Глава 2

4.1.1 Експериментални изследвания към комбинирания метод за разпознаване на активност чрез CNN и SVM

Първата част от експериментите се извършва върху PKU-MMD [47]. Това множество данни се състои от 51 различни типа активност, извършени от 66 субекти и записани от 3 камери в 3 различни ракурса, с общ брой на екземплярите – 21545. Това множество се използва за откриване на оптималните хиперпараметри. Активностите, които се състоят от два индивида (взаимодействие), са пренебрегнати, което остава около 19000 екземпляри. Оценява се производителността на модела, и неговите различни версии. (Таблица 4.1)

Таблица 4.1: Точност на предложения метод върху PKU-MMD при различни версии на метода

Подход	Точност (%)
Само горен слой	73.6
Горен слой + SVM	82.6
И двата слоя	88.8
И двата слоя + SVM	**92.2**

Втората част от експериментите се извършва върху NTU RGB+D. Това е обемно множество данни, съдържащо записи от анотирани активности на индивиди. То се състои от 56880 екземпляра, обхващащи 60 различни класове активности, снимани от 3 различни ракурса (през 45 градуса) от 40 различни индивиди (на възраст между 10 и 35 години). Авторите на това множество предлагат две метрики за оценка (benchmark) на модел върху него – cross-subject (между различните субекти) и cross-view (между различните ракурси). 11 от класовете в NTU RGB+D са за взаимодействие между двама индивиди, например прегръщане и ръкостискане. За справянето с тези класове, в началото поотделно са взети техните матрици от скелетони, след което са конкатенирани вертикално и взети като един екземпляр. По-нататъшната обработка не се променя.

Резултатите са показани на Таблица 4.2, заедно с резултати от известни подходи в литературата.

Анализ на резултатите

Резултатите при различните конфигурации, че SVM е подходящ класификатор за допълнителното повишаване на класификационната способност на даден модел. На базата на постигнатата висока точност (над 90%) може да се установи, че CNN-SVM е подходящ, както за средни по мащаб множества данни като PKU-MMD, така и малки

Таблица 4.2: Таблица с резултати върху NTU-RGB+D dataset за подходи от литературата и представеният

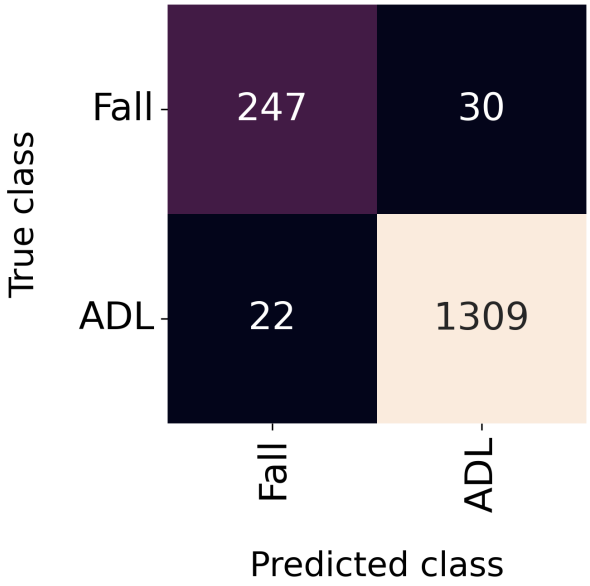
Подход	Точност (%) Cross-Subject	Точност (%) Cross-View
Deep RNN [2]	56.3	64.1
Part-aware [2] LSTM	62.9	70.3
CNN+LSTM [58]	67.5	76.2
ST-LSTM [95]	69.2	77.7
SkeleMotion [34]	**69.6**	80.1
Двупоточна DNN + SVM (предложен метод)	69.1	**81.1**

множества като UTD-MHAD (с под 1000 екземпляра). Включването на допълнителния слой за векторите от разлики допълнително повишава дискриминативността на CNN, което демонстрира наличие на нова и информативна перспектива върху активностите.

При NTU-RGB+D резултатите показват точност, надвишаваща тази на невронните мрежи с памет и сходна при подобни методи като [34], където има представяне на активността като изображение. Разликата в точността при cross-view benchmark-а спрямо cross-subject показва, че моделът е устойчив на промени в ракурса, което означава, че е подходящ за използване в многоракурсна система.

4.1.2 Експериментални изследвания към метода за откриване на необичайна човешка активност в реално време с помощта на 1DCNN-LSTM

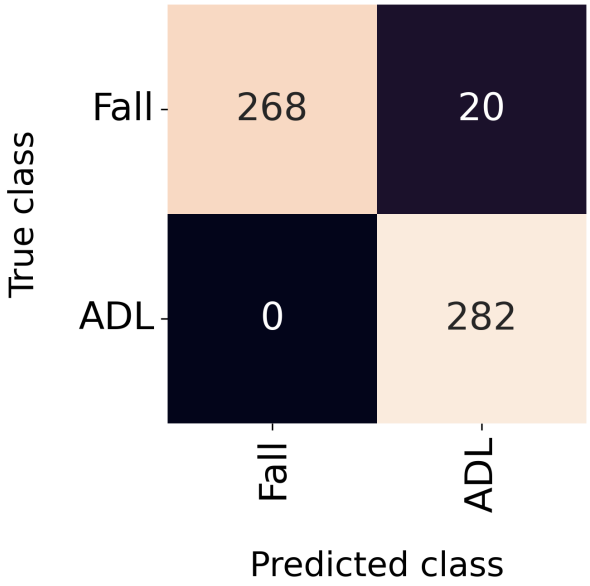
Извършена е класификация върху подмножество от NTU-RGB+D. Следните шест класа са избрани за този експеримент: ”вдигане”, ”сядане”, ”изправяне”, ”отговаряне на телефона”, ”ходене” и ”падане”. Всички те, с изключение на ”падане”, са групирани в един клас за дейности от ежедневието. Бенчмаркът за cross-subject е използван за избор на обучаващи и тестови подмножества, които се състоят съответно от 4005 и 1608 проби след селектирането. Моделът е обучен в 2 епохи. Неговата ефективност може да се види в матрицата на объркването (Фиг. 4.4)



Фигура 4.4: Матрица на объркване от тестовия набор на NTU-RGB+D.
Точност - 0.967

Като втори експеримент върху NTU-RGB+D се избират всичките класове, в които активностите включват един индивид. Разделят се на два класа ”падане” и всички останали класове, като от тези класове се избира точен брой екземпляри на случаен принцип, така че общият брой да е равен на този на клас ”падане”. Цялото множество се разделя в съотношение 7:3 обучение/тест. Това разделение е използвано и в статията [74], в която авторите изследват производителността на Spatio-temporal Graph convolutional network (ST-GCN).

Обучението се извършва в 2 епохи като предходния експеримент. Точността на 1DCNN-LSTM е представена във Фиг. 4.7. Точността на режата, както и нейни варианти само с 1DCNN и само с LSTM, е сравнена със ST-GCN от [74] в Таблица 4.3. В тази таблица са сравнени и броят параметри, които са пропорционални на скоростта за обучение и извеждане на данни на мрежите.



Фигура 4.7: Матрица на объркване от тестовия набор на балансираното подмножество от NTU-RGB+D. Точност - 0.965.

Таблица 4.3: Точност и брой параметри на предложения метод върху подмножеството NTU-RGB+D при различни конфигурации.

Подход	Точност (%)	Брой параметри
Само 1DCNN	96.1	86794
Само LSTM	95.2	46362
1DCNN-LSTM	**96.5**	133306
ST-GCN [74]	92.9	≈ 3096000 [38]

Анализ на резултатите

Моделът 1DCNN-LSTM демонстрира висока точност на класификация върху две различни подмножества от NTU-RGB+D (над 90%). Върху балансирано подмножество от NTU-RGB+D предложената архитектура надвишава откъм точност (96.5%) дълбока невронна мрежа ST-GCN, предложена за откриване на падане в [74] (92.9%).

4.1.3 Експериментални изследвания към метода за откриване на взаимодействие човек-обект

Наборът от данни, който се извлича от PKU-MMD, е разделен на части 70:20:10 съответно за обучение, валидиране и тестване. Всяка последователност е разделена на прозорци от 30 кадъра със стъпка 10. След това мрежата се обучава на прозорците за 10 епохи. Избраният оптимизатор е Adam, със скорост на обучение 0.01. Тъй като имаме строго небалансирано множество данни, където кадрите на взаимодействие са много по-малко, за функция на загубата се използва притеглена двоична крос-ентропия (weighted

binary cross-entropy), където теглото за етикет ”1” (наличие на взаимодействие) е 0.75, а за ”0” – 0.25. Accuracy, recall, precision и F1 резултат са избрани като показатели за ефективност и те са изчислени въз основа на прагово определяне на крайните резултати с различни стойности. Оптималните резултати се достигат с праг от 0.2. (Таблица. 4.4)

Таблица 4.4: Метрики получени от обучения модел при различни зададени прагове за двоична кросентропия

Threshold	Accuracy	Recall	Precision	F1 Score
0.5	0.922	0.627	0.875	0.731
0.35	0.928	0.669	0.878	0.759
0.25	0.934	0.708	0.879	0.784
0.2	**0.934**	**0.732**	**0.855**	**0.789**

Анализ на резултатите

Постигнатите точност, precision и recall от експериментите са високи (над 0.7). Сравнение с други научни изследвания е трудно поради спецификата на задачата и множеството данни. Най-сходното изследване е това на Bruckschen et. al [68], където precision е по-висок (0.82 срещу 0.855), докато recall е по-нисък (0.82 срещу 0.732). При най-ниския тестван праг от 0.2 се наблюдават най-високи резултати, както е открито и в [68] (праг 0.22).

4.2 Експериментални изследвания на методите към Глава 3

4.2.1 Сливане чрез оптимална поза и класификация на PKU-MMD чрез скрити Марковски модели

Множеството данни PKU-MMD е избрано за експериментите към този метод поради заснемането му в многоракурсна обстановка, както и по-интензивното движение на индивидите около сцената. Скелетонните последователности, заснети от 3-те ракурса, са етикирани от различни хора и не са синхронизирани.

Броят на скритите състояния в Марковските модели е избран от 7 експеримента върху първоначалното необединено множество данни като броят се променя из стойностите - 5, 10, 15, 20, 25, 30, 35. Разликата в точността е показана в Таблица 4.5.

Таблица 4.5: Начални класификационни резултати за PKU-MMD чрез HMM при различни стойности на броя скрити състояния

Брой скрити състояния	5	10	15	20	25	30	35
Accuracy	0.586	0.618	0.662	0.667	0.685	0.677	0.680
Mean Recall	0.593	0.628	0.669	0.679	0.700	0.690	0.694
Mean Precision	0.604	0.627	0.668	0.679	0.693	0.689	0.692

От Таблица 4.5 се вижда, че най-високите резултати се достигат при около 20 скрити състояния, след което няма значително нарастване. Затова 20 е броят на скритите състояния, при които е направена класификация върху множеството, преди и след сливане. Точността на множество преди сливане е **0.667**, а след - **0.720**.

4.2.2 Сливане чрез филтър на Калман и класификация на NTU-RGB чрез HCN

При NTU-RGB почти всяка активност е заснета от 3 сензора Kinect с 45 градусово отместване от сцената. За експеримента са взети предвид активностите, които имат точно 3 заснемания от 3-те ракурса. Разделението обучение-тест се прави спрямо cross-subject разделението на NTU-RGB. След филтриране оставащите данни за обучение са 33162 на брой, а тестовите са 11028. Тъй като HCN приема активности с константен размер, последователностите се мащабират по оста на времето чрез билинейна интерполация, която се използва от авторите на мрежата. Размерите, до които те мащабират входните данни, са 32 и 64, при които авторите получават най-висока точност. В контекста на този експеримент избраният размер е 96, тъй като е максимално близо до средната продължителност на една активност в NTU-RGB (85), а прекалено малък зададен размер ще премахне детайлните междукадрови разлики, които са нужни за разграничаването между сходните класове активности. Авторите обучават HCN за 400 епохи със скорост на обучение 0.001, която намалява експоненциално при продължително обучение - на всяка епоха текущата скорост се умножава с коефициент от 0.999. Размерът на batch-овете е 64. Поради по-малкия обем на подмножеството избрани данни, броят на епохите е избран да бъде 100, тъй като тестовата точност не показва покачване към края на този период.

В Таблица 4.9 са изведени сумарните резултати преди и след сливане, както и тези при обучението и тестването на всеки ракурс поотделно.

Таблица 4.9: Точност, precision и recall при различни конфигурации на NTU-RGB+D

Конфигурация	Precision	Recall	Точност
Преди сливане (всички данни)	0.826	0.831	0.827
Само ляв ракурс	0.759	0.747	0.747
Само среден ракурс	0.749	0.747	0.748
Само десен ракурс	0.752	0.747	0.747
Сливане чрез филтър на Калман	0.797	0.801	0.797

Анализ на резултатите

Повишаването на точността на класификация (от 66.7% на 72%) при скритите Марковски модели показва, че при тях сливането чрез оптимална поза е подходящ подход, тъй като те не взимат предвид междукадровите разлики. По-ниската класификационна точност на HMM класификатора спрямо дълбоките невронни мрежи като например CNN в в 4.1.1 (72% спрямо 73.6%) показва, че те не са подходящи за по-големи множества от данни, каквото е PKU-MMD. Това се дължи на факта, че покласово те не

споделят параметри помежду си (всеки модел е независим един от друг), откъдето идва объркването между подобни класове активност. Сливането чрез филтър на Калман не подобрява класификацията при NTU-RGB+D, а намаля точността с около 3%. При сравнение с обучението върху който и да е ракурс от NTU-RGB+D поотделно се наблюдава повишаване на точността с около 5%.

III. ПРИНОСИ В ДИСЕРТАЦИОННИЯ ТРУД

Научните приноси са:

1. Разработен е комбиниран, каскаден метод за разпознаване на човешка активност, включващ конволюционна невронна мрежа (CNN) и машина за опорни вектори (SVM). Постигнати са точности от 69.1% и 81.1% при CS и CV на NTU-RGB+D. Резултатите надвишават тези на [58](2% CS и 5% CV подобрение). Използването на SVM позволява по-висока точност при по-кратко обучение и по-малко количество данни.

   Приносът е към Глава 2, точка 2.2.

Научно-приложните приноси са:

1. Разработена е архитектура на комбинирана невронна мрежа за откриване на аномална човешка активност, включващ едномерна конволюционна невронна мрежа (1DCNN) и мрежа с дълга краткосрочна памет (LSTM). Постигната е точност от 96.5% върху балансирано подмножество от NTU-RGB+D. Резултатът надвишава този, предложен в [74] (92.9%).

   Приносът е към Глава 2, точка 2.3.

2. Разработен е алгоритъм за покадрова класификация на активности за наличие на взаимодействие между индивид и обект. Постигната е класификационна покадрова точност от 93.4%, recall - 73.2% и precision 85.5% при праг на взаимодействие 0.2. Тези резултати са представени в [D4].

   Приносът е към Глава 2, точка 2.4.

3. Разработен е алгоритъм за разпознаване на многоракурсни активности от скелетони чрез сливане на данните. Постигната е класификационна точност от 66.7% преди сливане на данните и 72% след сливане на данните при класификатор от множество Марковски модели.

   Приносът е към Глава 3, точка 3.3.

Приложните приноси са:

1. Експериментални изследвания на представените методи и алгоритми.

   Този принос е към Глава 4, точки 4.1.1, 4.1.2, 4.1.3, 4.2.1 и 4.2.2.

IV. ЗАКЛЮЧЕНИЕ

Дълбоките невронни мрежи подобряват качествено анализа на активност и превъзхождат класическите модели в задачите, представени в тази дисертация. Независимо от това, времето им за обучение нараства със сложността на задачата и обема данни. Сливането на данните от много ракурси е едно от решенията на този проблем. Методът с оптимална поза показва приблизителни класификационни резултати на тези, където се използват всички данни. Независимо от това, този тип методи изискват непрекъснатото проследяване от няколко заснемащи устройства във всеки един момент от времето. При сценарий, в който това условие не е спазено, производителността ще спадне при тестване върху неоптимизираните данни. За щастие, в последните години се набляга на темата за one-shot обучението, при което дълбоките мрежи се опитват да научат всичко за данните от един до няколко екземпляра.

V. ИЗПОЛЗВАНИ СЪКРАЩЕНИЯ

CMOS **C**omplementary **m**etal-**o**xide **s**emiconductor
STIP **S**patio-**T**emporal **I**nterest **P**oints
IoT **I**nternet-**o**f-**T**hings
moCap **m**otion **C**apture
HAR **H**uman **A**ctivity **R**ecognition
HOI **H**uman **O**bject **I**nteraction
3D **T**hree **d**imensional
ToF **T**ime **o**f **F**light
SDK **S**oftware **D**evelopment **K**it
RGBA **R**ed **G**reen **B**lue **A**lpha
RGBD **R**ed **G**reen **B**lue + **D**epth
GPU **G**raphics **P**rocessing **U**nit
DNN **D**eep **N**eural **N**etwork
CNN **C**onvolutional **N**eural **N**etwork
1D CNN **O**(ne) **D**imensional **C**NN
SVM **S**upport **V**ector **M**achine
LSTM **L**ong **S**hort **T**erm **M**emory
YOLO **Y**ou **O**nly **L**ook **O**nce
CSR-DCF **D**iscriminative **C**orrelation **F**ilter with **C**hannel and **S**patial **R**eliability
HMM **H**idden **M**arkov **M**odel
DTW **D**ynamic **T**ime **W**arping
HCN **H**ierarchal **C**o-ocurrence **N**etwork
PKU-MMD **P**eking **U**niversity **M**ulti-**M**odal **D**ataset
NTU-RGB+D **N**anyang **T**echnological **U**niversity **R**ed **G**reen **B**lue + **D**epth dataset
CS **C**ross-**s**ubject
CV **C**ross-**v**iew

VI. СПИСЪК С ПУБЛИКАЦИИТЕ ПО ДИСЕРТАЦИЯТА

[D1] P. Hristov, P. Nikolov, A. Manolova, and O. Boumbarov, “Multi-view RGB-D System for Person Specific Activity Recognition in the context of holographic communication”, in *2020 XXIX International Scientific Conference Electronics (ET)*, 2020, pp. 1–4. DOI: 10.1109/ET50336.2020.9238233

[D2] P. Hristov, A. Manolova, and O. Boumbarov, "Deep Learning and SVM-Based Method for Human Activity Recognition with Skeleton Data", in *2020 28th National Conference with International Participation (TELECOM)*, 2020, pp. 49–52. DOI: 10.1109/TELECOM50385.2020.9299541

[D3] P. Hristov, "Real-time Abnormal Human Activity Detection Using 1DCNN-LSTM for 3D Skeleton Data", in *2021 12th National Conference with International Participation (ELECTRONICA)*, 2021. DOI: 10.1109/ELECTRONICA52725.2021.9513696

[D4] P. Hristov, D. Avresky, and O. Boumbarov, "Human-Object Interaction Detection: 1D Convolutional Neural Network Approach Using Skeleton Data", in *2021 IEEE 20th International Symposium on Network Computing and Applications (NCA)*, 2021, pp. 1–5. DOI: 10.1109/NCA53618.2021.9685549

VII. ИЗПОЛЗВАНА ЛИТЕРАТУРА

[1] L. Chunhui, H. Yueyu, L. Yanghao, S. Sijie, and L. Jiaying, "Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding", *arXiv preprint arXiv:1703.07475*, 2017

[2] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019

[3] W. Kay, J. Carreira, K. Simonyan, *et al.*, *The kinetics human action video dataset*, 2017. arXiv: 1705.06950 [cs.CV]

[4] Y. Liu, L. Nie, L. Liu, and D. S. Rosenblum, "From action to activity: Sensor-based activity recognition", *Neurocomputing*, vol. 181, pp. 108–115, 2016, ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2015.08.096

[5] M. G. Morshed, T. Sultana, A. Alam, and Y.-K. Lee, "Human action recognition: A taxonomy-based survey, updates, and opportunities", *Sensors*, vol. 23, no. 4, 2023, ISSN: 1424-8220. DOI: 10.3390/s23042182

[34] C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos, and W. R. Schwartz, "Skelemotion: A new representation of skeleton joint sequences based on motion information for 3d action recognition", in *2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2019, pp. 1–8. DOI: 10.1109/AVSS.2019.8909840

[38] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition", in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, 2018, ISBN: 978-1-57735-800-8

[42] X. Wang, J. Ellul, and G. Azzopardi, "Elderly fall detection systems: A literature survey", *Frontiers in Robotics and AI*, vol. 7, 2020

[48] L. Minh Dang, K. Min, H. Wang, M. Jalil Piran, C. Hee Lee, and H. Moon, "Sensor-based and vision-based human activity recognition: A comprehensive survey", *Pattern Recognition*, vol. 108, p. 107 561, 2020. DOI: https://doi.org/10.1016/j.patcog.2020.107561

[58] J. C. Núñez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Vélez, "Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition", *Pattern Recognition*, vol. 76, pp. 80–94, 2018, ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2017.10.033

[68] L. Bruckschen, S. Amft, J. Tanke, J. Gall, and M. Bennewitz, "Detection of generic human-object interactions in video streams", in *Social Robotics*, 2019, pp. 108–118, ISBN: 978-3-030-35888-4

[74] O. Keskes and R. Noumeir, "Vision-based fall detection using st-gcn", *IEEE Access*, vol. 9, pp. 28 224–28 236, 2021. DOI: 10.1109/ACCESS.2021.3058219

[90] M. Edwards and R. Green, "Low-latency filtering of kinect skeleton data for video game control", in *Proceedings of the 29th International Conference on Image and Vision Computing New Zealand*, 2014, pp. 190–195, ISBN: 9781450331845. DOI: 10.1145/2683405.2683453

[91] D. Willner, C. B. Chang, and K. P. Dunn, "Kalman filter algorithms for a multi-sensor system", in *1976 IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes*, 1976, pp. 570–574. DOI: 10.1109/CDC.1976.267794

[92] W. Kabsch, "A solution for the best rotation to relate two sets of vectors", *Acta Crystallographica Section A*, vol. 32, pp. 922–923, 1976. [Online]. Available: https://api.semanticscholar.org/CorpusID:97383637

[95] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition", vol. 9907, Oct. 2016. DOI: 10.1007/978-3-319-46487-9_50

TECHNICAL UNIVERSITY OF SOFIA
FACULTY OF TELECOMMUNICATIONS
DEPARTMENT OF RADIOCOMMUNICATIONS
AND VIDEOTECHNOLOGIES

Plamen Atanasov Hristov, M.Sc.

VISUAL ANALYSIS OF THE BEHAVIOR OF INDIVIDUALS IN CYBER-PHYSICAL SYSTEMS

ABSTRACT of Ph.D. THESIS

The main topics of this PhD thesis are the development and application of methods and algorithms for human activity analysis, based on a skeleton-pose representation. Three methods for human activity analysis are proposed, two of which for multi-view systems.

The first method introduces a combined classifier, consisting of a Convolutional Neural Network (CNN) for feature extraction and a Support Vector Machine (SVM) for classification. The CNN serves as a separate classifier which is trained initially on full datasets of skeleton activities. It receives two types of activity representation as its input - the first being stacked matrices of the activities, one matrix for each joint coordinate. The second is a vector of frame-wise distances. The accuracy is improved when using the full method, compared to only using CNN or only one of its two inputs.

The second method combines two neural networks, which are purposed for time-series analysis - 1DCNN and a Long Short-Term Memory network (LSTM). The combined neural network 1DCNN-LSTM is trained on two-class datasets for fall detection where the positive class is the "falling" action.

The third method consists of a dataset creation and annotation step for human-object interaction, which is based on 2D video data and skeleton data. Windows around the skeletons are extracted and used for training an object-recognition deep neural network - YOLOv5. In the testing phase those windows are used for finding and tracking objects. The rules for human-object interaction are based on the movement of the object as it is assumed to be moved by a human. The coordinates of the object and the joints of the skeleton arm are used as frame-vectors for training a 1DCNN, which learns to generalize interaction from the strictly defined rules.

The next two methods are multi-view fusion methods, which are based on an optimal pose selection from a set of views. An optimal pose is that is which is facing forward towards the recording sensor or camera. The optimal pose is assumed to have the least amount of noise and error in joint calculation.

The purpose of the PhD thesis is to research and propose novel algorithms for human activity analysis in mono-view and multi-view systems, using deep neural networks. To study their performance, the algorithms are evaluated using popular and large-scale datasets of human activities.