

К. Ратгеб | Р. Толосана | Р. Вера-Родригес | К. Буш



Цифровые манипуляции с лицами и их обнаружение

От дипфейков до морфинг-атак

Содержание

От издательства	16
Предисловие	17
Часть I. ВВЕДЕНИЕ	19
Глава 1. Введение в цифровые манипуляции с лицами	20
1.1. Введение	21
1.2. Типы цифровых манипуляций с лицами	23
1.2.1. Синтез целевого лица	23
1.2.2. Замена идентичности	26
1.2.3. Морфинг лица	32
1.2.4. Манипуляции с характерными признаками лиц	33
1.2.5. Изменение выражения лица	35
1.2.6. «Аудио в видео» и «текст в видео»	37
1.3. Выводы	39
Литература	40
Глава 2. Цифровые манипуляции с лицами в биометрических системах	46
2.1. Введение	47
2.2. Биометрические системы	48
2.2.1. Процессы	49
2.2.2. Распознавание лиц	50
2.3. Цифровые манипуляции с лицами в биометрических системах	51
2.3.1. Влияние на биометрические характеристики	51
2.3.2. Методы обнаружения манипуляций	53
2.4. Эксперименты	56
2.4.1. Постановка эксперимента	56
2.4.2. Оценка эффективности	58
2.5. Выводы и перспективы	61
Литература	62
Глава 3. Мультимедийная криминалистика до эпохи глубокого обучения	65
3.1. Введение	66
3.2. Метод на основе PRNU	68

3.2.1. Определение PRNU	70
3.2.2. Вычисление остаточного шума	71
3.2.3. Тест на обнаружение подделки.....	71
3.2.4. Анализ на основе управляемой фильтрации	73
3.3. Слепые методы	76
3.3.1. Паттерны шума	76
3.3.2. Артефакты компрессии	80
3.3.3. Артефакты редактирования.....	82
3.4. Методы обучения с признаками, созданными вручную.....	84
3.5. Выводы.....	85
Литература	87

Часть II. ЦИФРОВЫЕ МАНИПУЛЯЦИИ С ЛИЦАМИ И ПРИЛОЖЕНИЯ БЕЗОПАСНОСТИ

Глава 4. Создание дипфейков и борьба с ними.....	92
4.1. Введение	92
4.2. Основы	96
4.2.1. Генерация дипфейк-видео	96
4.2.2. Методы обнаружения дипфейков	97
4.2.3. Существующие наборы данных дипфейков	98
4.3. Celeb-DF: создание дипфейков	99
4.3.1. Метод синтеза	100
4.3.2. Визуальное качество	103
4.3.3. Оценки.....	104
4.4. Landmark Breaker: препятствие для DeepFake	106
4.4.1. Экстракторы лицевых отметок.....	106
4.4.2. Состязательные возмущения.....	106
4.4.3. Обозначения и формулировка.....	107
4.4.4. Оптимизация.....	107
4.4.5. Установки эксперимента	108
4.4.6. Результаты	110
4.4.7. Анализ устойчивости.....	112
4.4.8. Исследование абляции	114
4.5. Заключение	115
Литература	116

Глава 5. Угроза дипфейков для компьютерного зрения и человеческого зрительного восприятия

5.1. Введение	119
5.2. Сопутствующие работы.....	121
5.3. Базы данных и методы	122
5.3.1. DeepfakeTIMIT	122
5.3.2. DF-Mobio	123
5.3.3. Google и Jigsaw.....	124
5.3.4. Facebook.....	124

5.3.5. Celeb-DF	125
5.4. Протоколы оценки эффективности	126
5.4.1. Измерение уязвимости	126
5.4.2. Измерение эффективности распознавания дипфейков	127
5.5. Уязвимость систем распознавания лиц	127
5.6. Субъективная оценка человеческого визуального восприятия	128
5.6.1. Результаты субъективной оценки	131
5.7. Оценка алгоритмов обнаружения дипфейков	134
5.8. Заключение	136
Литература	137

Глава 6. Создание морфа и уязвимость систем распознавания

лиц к морфингу	139
6.1. Введение	140
6.2. Генерация морфинга лица	143
6.2.1. Морфинг на основе лицевых отметок	143
6.2.2. Генерация морфинга лица на основе глубокого обучения	149
6.3. Уязвимость систем распознавания лиц к морфированию лица	151
6.3.1. Наборы данных	152
6.3.2. Результаты	153
6.3.3. Результаты морфинга на основе глубокого обучения	158
6.4. Выводы	158
Литература	159

Глава 7. Состязательные атаки на системы распознавания лиц

7.1. Введение	162
7.2. Классификация атак на FRS	165
7.2.1. Модель угрозы	166
7.3. Отравляющие атаки на FRS	170
7.3.1. Метод быстрого градиентного знака	170
7.3.2. Прогнозируемый градиентный спуск	170
7.4. Атаки Карлини и Вагнера (CW)	171
7.5. Модель ArcFace FRS	172
7.6. Эксперименты и анализ	173
7.6.1. Чистый набор данных	173
7.6.2. Набор данных атак	174
7.6.3. Модель FRS для базовой проверки	175
7.6.4. Базовая оценка эффективности FRS	175
7.6.5. Эффективность FRS при отравлении проверочных данных	178
7.6.6. Эффективность FRS при отравлении данных регистрации	179
7.7. Столкновение состязательного обучения с атаками FGSM	180
7.8. Обсуждение	182
7.9. Выводы и будущие направления разработок	183
Литература	183

Глава 8. Генерация говорящих лиц: «аудио в видео»	187
8.1. Введение	187
8.2. Сопутствующие методы	189
8.2.1. Звуковое представление	189
8.2.2. Моделирование лица	190
8.2.3. Анимация звук–лицо	194
8.2.4. Постпроцессинг	201
8.3. Наборы данных и метрики	201
8.3.1. Набор данных	201
8.3.2. Метрики	203
8.4. Обсуждение	205
8.4.1. Тонкий контроль лица	205
8.4.2. Обобщение	207
8.5. Заключение	208
8.6. Дополнительная литература	209
Литература	209

Часть III. ОБНАРУЖЕНИЕ ЦИФРОВЫХ МАНИПУЛЯЦИЙ С ЛИЦАМИ

216

Глава 9. Обнаружение синтетических лиц, созданных искусственным интеллектом	217
9.1. Введение	218
9.2. Генерация лиц с помощью искусственного интеллекта	220
9.3. Отпечатки пальцев GAN	221
9.4. Методы обнаружения в пространственной области	224
9.4.1. Признаки ручной работы	225
9.4.2. Признаки, управляемые данными	226
9.5. Методы обнаружения по областям частот	227
9.6. Обучение обобщающих особенностей	228
9.7. Обобщающий анализ	230
9.8. Анализ надежности	232
9.9. Дальнейший анализ обнаружения GAN	233
9.10. Нерешенные проблемы	235
Литература	238

Глава 10. 3D-архитектура CNN и механизмы внимания для обнаружения дипфейков	242
10.1. Введение	243
10.2. Сопутствующие исследования	245
10.2.1. Обнаружение дипфейков	245
10.2.2. Механизмы внимания	247
10.3. Набор данных	253
10.4. Алгоритмы	254
10.5. Эксперименты	254

10.5.1. Все техники манипуляции.....	255
10.5.2. Отдельные техники манипуляций	257
10.5.3. Техники перекрестной манипуляции.....	258
10.5.4. Эффект внимания в 3D ResNets	259
10.5.5. Визуализация соответствующих признаков в обнаружении дипфейка.....	260
10.6. Выводы.....	260
Литература	262

Глава 11. Обнаружение дипфейков с использованием

нескольких модальностей данных	266
11.1. Введение	267
11.2. Обнаружение дипфейков с помощью пространственно-временных особенностей видео	268
11.2.1. Обзор	270
11.2.2. Модельный компонент	270
11.2.3. Детали обучения	273
11.2.4. Бустинговая нейронная сеть	273
11.2.5. Аугментация времени тестирования.....	274
11.2.6. Анализ результатов	274
11.3. Обнаружение дипфейков с помощью анализа аудиоспектрограммы....	276
11.3.1. Обзор	277
11.3.2. Набор данных	278
11.3.3. Генерация спектрограммы.....	278
11.3.4. Сверточная нейронная сеть (CNN)	279
11.3.5. Результаты экспериментов	280
11.4. Обнаружение дипфейков посредством анализа несоответствия аудио и видео.....	280
11.4.1. Обнаружение несоответствия аудио и видео посредством несоответствия фонем и визем	282
11.4.2. Обнаружение дипфейков с использованием аффективных сигналов	284
11.5. Заключение	286
Литература	286

Глава 12. Обнаружение дипфейков на основе определения

сердечного ритма: однокадровый и многокадровый методы	289
12.1. Введение	290
12.2. Сопутствующие работы.....	293
12.3. DeepFakesON-Phys	297
12.4. Базы данных.....	298
12.4.1. База данных Celeb-DF v2.....	298
12.4.2. DFDC Preview	299
12.5. Экспериментальный протокол	299
12.6. Результаты обнаружения фейков: DeepFakesON-Phys	300

12.6.1. Обнаружение дипфейков на уровне кадра	300
12.6.2. Обнаружение дипфейков на уровне короткого видео	303
12.7. Выводы	304
Литература	306

Глава 13. Капсульно-криминалистические сети

для обнаружения дипфейков	310
13.1. Введение	311
13.2. Сопутствующие работы	313
13.2.1. Генерация дипфейков	313
13.2.2. Обнаружение дипфейков	314
13.2.3. Проблемы обнаружения дипфейков	315
13.2.4. Капсульные сети	316
13.3. Капсульная криминалистика	316
13.3.1. Зачем нужна капсульная криминалистика?	316
13.3.2. Обзор	317
13.3.3. Архитектура	317
13.3.4. Алгоритм динамической маршрутизации	319
13.3.5. Визуализация	321
13.4. Оценка	321
13.4.1. Наборы данных	324
13.4.2. Метрики	325
13.4.3. Эффект улучшений	325
13.4.4. Сравнение экстракторов особенностей лиц	326
13.4.5. Влияние слоев статистического пулинга	327
13.4.6. Сеть Capsule-Forensics по сравнению с CNN: замеченные атаки	328
13.4.7. Сеть Capsule-Forensics против CNN: невидимые атаки	330
13.5. Заключение и будущая работа	333
13.6. Приложение	333
Литература	335

Глава 14. Обнаружение дипфейков: набор данных

DeeperForensics и постановка задачи	339
14.1. Введение	340
14.2. Сопутствующие работы	342
14.2.1. Методы создания дипфейков	342
14.2.2. Методы обнаружения дипфейков	343
14.2.3. Наборы данных для обнаружения дипфейков	344
14.2.4. Лучшие тесты обнаружения дипфейков	345
14.3. Набор данных DeeperForensics-1.0	346
14.3.1. Сбор данных	346
14.3.2. Вариационный автокодировщик дипфейков	348
14.3.3. Масштаб и разнообразие	353
14.3.4. Набор скрытых тестов	354
14.4. DeeperForensics Challenge 2020	355

14.4.1. Платформа	356
14.4.2. Набор данных задачи	356
14.4.3. Критерии оценки	356
14.4.4. Таймлайн	357
14.4.5. Результаты и решения	357
14.5. Обсуждение	362
14.6. Дополнительная литература	362
Литература	363

Глава 15. Методы обнаружения морфинговых атак лица 369

15.1. Введение	369
15.2. Сопутствующие работы	371
15.3. Конвейер обнаружения морфинговых атак	372
15.3.1. Подготовка данных и извлечение признаков	373
15.3.2. Подготовка признаков и обучение классификатора	373
15.4. База данных	374
15.4.1. Морфинг изображения	376
15.4.2. Постпроцессинг изображения	378
15.5. Методы обнаружения морфинговых атак	379
15.5.1. Предварительная обработка	379
15.5.2. Извлечение признаков	380
15.5.3. Классификация	382
15.6. Эксперименты	382
15.6.1. Обобщаемость	383
15.6.2. Эффективность обнаружения	384
15.6.3. Постпроцессинг	384
15.7. Заключение	386
Литература	387

Глава 16. Практическая оценка методов обнаружения морфинговых атак лица 390

16.1. Введение	391
16.2. Сопутствующие работы	393
16.3. Создание наборов данных морфинга	394
16.3.1. Создание морфов	394
16.3.2. Наборы данных	395
16.4. Обнаружение морфинговых атак лиц на основе текстур	396
16.5. Маскировка морфинга	397
16.6. Эксперименты и результаты	399
16.6.1. Эффективность набора данных	399
16.6.2. Эффективность перекрестного набора данных	400
16.6.3. Эффективность смешанного набора данных	400
16.6.4. Устойчивость к аддитивному гауссову шуму	401
16.6.5. Устойчивость к масштабированию	401
16.6.6. Выбор субъектов с похожими лицами	402
16.7. Детектор SOTAMD	403

16.8. Заключение	404
Литература	404

Глава 17. Ретушь лица и обнаружение изменений.....407

17.1. Введение	408
17.2. Ретуширование и обнаружение изменений – обзор.....	410
17.2.1. Обнаружение цифровой ретуши	410
17.2.2. Обнаружение цифровых изменений	413
17.2.3. Общедоступные базы данных	415
17.3. Экспериментальная оценка и наблюдения	417
17.3.1. Обнаружение междоменных изменений	420
17.3.2. Обнаружение изменений перекрестных манипуляций.....	421
17.3.3. Обнаружение межэтнических изменений	422
17.4. Нерешенные проблемы	423
17.5. Заключение.....	424
Литература	425

Часть IV. ДАЛЬНЕЙШИЕ ТЕМЫ, ТЕНДЕНЦИИ И ПРОБЛЕМЫ.....429

Глава 18. Улучшение конфиденциальности мягкой биометрии.....430

18.1. Введение	431
18.2. Предыстория и сопутствующие работы	434
18.2.1. Формулировка проблемы и существующие решения.....	434
18.2.2. Модели мягкобиометрической конфиденциальности	435
18.2.3. Обнаружение повышения конфиденциальности	437
18.3. Обнаружение вмешательства через несоответствие прогнозов (PREM).....	437
18.3.1. Обзор PREM	438
18.3.2. Сверхвысокое разрешение для восстановления признаков	439
18.3.3. Измерение несоответствия прогноза.....	440
18.3.4. Краткое описание и характеристики PREM.....	441
18.4. Эксперименты и результаты	442
18.4.1. Наборы данных и экспериментальные установки	442
18.4.2. Используемые модели конфиденциальности	443
18.4.3. Детали реализации	444
18.4.4. Результаты и обсуждения	445
18.5. Заключение	450
Литература	451

Глава 19. Обнаружение манипуляций с лицами в удаленных операционных системах.....455

19.1. Введение	456
19.2. Удаленная регистрация документов, удостоверяющих личность	457
19.3. Алгоритмы манипуляции с лицом	458
19.3.1. Категории атак	458
19.3.2. Общие алгоритмы манипуляции с лицом	462

19.4. Обнаружение манипуляций с лицами	464
19.4.1. Методы, специфичные для лица	465
19.4.2. Методы, независимые от лица	466
19.4.3. Наборы данных	469
19.5. Контркриминалистика и меры противодействия	471
19.5.1. Контркриминалистика	471
19.5.2. Меры противодействия	472
19.6. Базовая структура, стандартизация и правовые аспекты	475
19.7. Выводы	476
Литература	477

Глава 20. Перспективы, социальные и этические проблемы, связанные с биометрией при удаленной адаптации

20.1. Введение	483
20.2. Похищение идентичности и растущая потребность в ее удаленной проверке	485
20.2.1. Риски и социальные последствия похищения идентичности	485
20.2.2. Необходимость удаленной биометрической верификации идентичности	487
20.3. Технологии удаленной биометрической идентификации	490
20.3.1. Появление биометрической удаленной идентификации	490
20.3.2. Технологии удаленной биометрической идентификации	494
20.4. Этика, конфиденциальность и социальная приемлемость биометрической идентификации	497
20.4.1. Риски и основные этические проблемы	497
20.4.2. Целостность практической идентичности	500
20.4.3. Конфиденциальность и функциональные нарушения	501
20.4.4. Этические проблемы, возникающие в результате алгоритмически обусловленных действий и решений	503
20.4.5. Общественное признание технологии	505
20.5. Обсуждение и выводы	506
Литература	509

Глава 21. Грядущие тенденции в области цифровых манипуляций с лицами и их обнаружения

21.1. Введение	514
21.2. Реализм манипуляций с лицами и базы данных	515
21.2.1. Современное состояние	515
21.2.2. Недостающие ресурсы	517
21.3. Ограничения обнаружения манипуляций с лицами	518
21.3.1. Обобщаемость	518
21.3.2. Интерпретируемость	519
21.3.3. Слабые места детекторов	520
21.3.4. Возможности человека	520
21.3.5. Дальнейшие ограничения	521

21.4. Манипуляции с лицами и их обнаружение: путь вперед	521
21.4.1. Области применения манипуляций с лицами	521
21.4.2. Перспективные методы	524
21.5. Социальные и правовые аспекты манипуляции лицами и их обнаружения.....	525
21.6. Выводы.....	528
Литература	529
Предметный указатель.....	534