УДК 004.93

Свёрточные нейронные сети в системах обнаружения и распознавания лиц

В. В. Кочетуров, О. И. Федяев
Донецкий национальный технический университет
кафедра программной инженерии им. Л.П. Фельдмана
E-mail: v.v_kocheturov@mail.ru

Аннотация:
В статье рассматривается применение свёрточных нейронных сетей на различных этапах систем распознавания лиц: обнаружение, выравнивание и распознавание. Обозреваются ключевые архитектуры и методы, применяемые на каждом из этапов. Описываются многозадачные CNN, выполняющие несколько функций одновременно. Выделяются современные тенденции развития в этой области. Описываются многозадачные CNN, выполняющие несколько функций одновременно. Выделяются современные тенденции развития: применение механизмов внимания, трансформеров и обучение на малых данных.

Введение

Системы распознавания лиц стали неотъемлемой частью многих современных технологий, от биометрической аутентификации [1] до анализа видеопотоков в системах безопасности. В основе успеха этих систем лежат свёрточные нейронные сети (CNN), которые продемонстрировали выдающуюся способность к извлечению информативных признаков из изображений.

Процесс распознавания лиц является многоэтапным, и для каждого этапа существуют специализированные или адаптированные архитектуры CNN. Понимание особенностей этих архитектур и методов их применения критически важно для разработки эффективных и робастных систем. В данной статье рассматриваются современные CNN, используемые на этапах обнаружения, выравнивания, извлечения признаков и принятия решения, а также архитектуры, объединяющие некоторые из этих этапов.

Обнаружение лиц с использованием CNN

Обнаружение лиц – первый и один из важнейших этапов, заключающийся в локализации всех лиц на изображении или в видеопотоке. Детекторы лиц должны быть устойчивы к изменениям масштаба, ракурса, освещения, частичным окклюзиям и выражениям лиц. Высокая скорость работы также является критическим требованием для многих приложений.

Современные CNN-детекторы можно условно разделить на два основных типа (рис. 1).

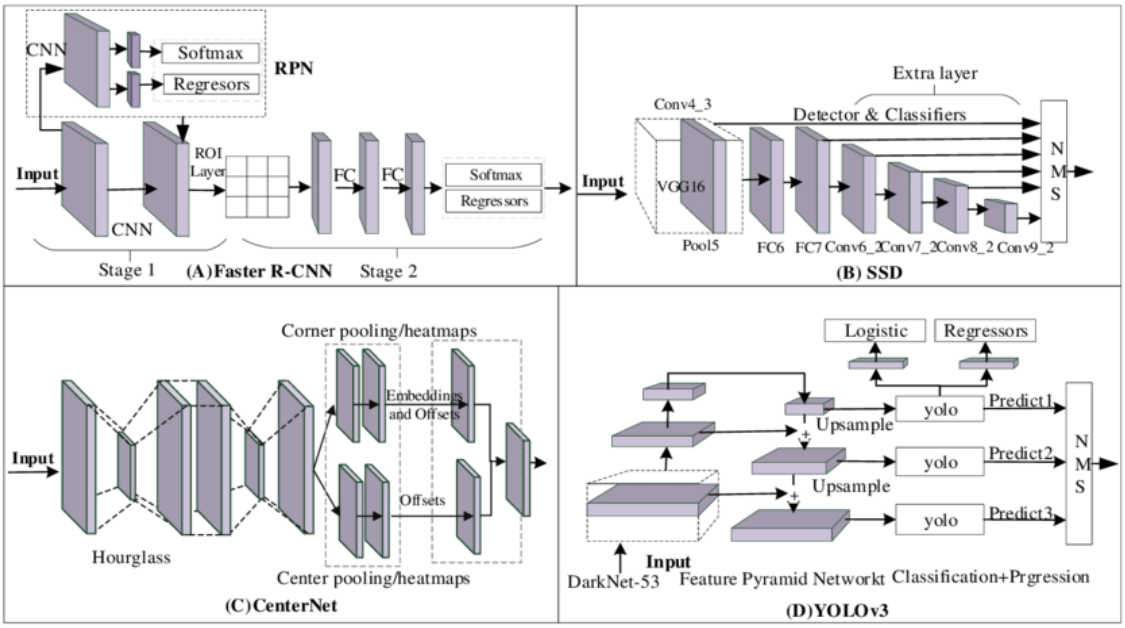

Рисунок 1 – Архитектуры CNN-детекторов (Faster R-CNN, SSD, CenterNet, YOLOv3)

Двухэтапные, такие как Faster R-CNN, сначала генерируют регионы-кандидаты, а затем классифицируют их [2]. Они обычно более точные, но медленнее. Одноэтапные, например, YOLO и SSD, предсказывают ограничивающие рамки и классы объектов за один проход по сети [3]. Они, как правило, быстрее, что делает их предпочтительными для задач реального времени. К одноэтапным также относится CenterNet, который использует подход "объекты как точки": вместо рамок он определяет центры объектов на тепловой карте и регрессирует их размеры. Это позволяет достичь хорошего баланса между точностью и скоростью, особенно в условиях отсутствия якорей.

Ключевыми архитектурами CNN для обнаружения являются:

– SSD (Single Shot MultiBox Detector) и YOLO (You Only Look Once): хотя это детекторы общего назначения, они успешно адаптируются для обнаружения лиц путем обучения на соответствующих датасетах. Они используют предсказания на картах признаков разного масштаба для детекции объектов разного размера [3].

– MTCNN (Multi-task Cascaded Convolutional Networks): Каскад из трёх сетей (P-Net, R-Net, O-Net), где первые сети (P-Net, R-Net) выполняют грубую детекцию и отсев кандидатов [4]. Подробнее будет рассмотрен в разделе интегрированных архитектур.

– RetinaFace: Одноэтапный детектор, часто использующий ResNet или MobileNet в качестве базовой сети (backbone) и Feature Pyramid Network (FPN) для эффективной работы с лицами разных масштабов [5].

Для увеличения поля обзора без увеличения числа параметров и сохранения разрешения карт признаков в детекторах лиц активно применяются расширенные свёртки (atrous/dilated convolutions). Это позволяет сетям захватывать более широкий контекст, что особенно важно для различения лиц на сложном фоне или для анализа связей между частями лица. Например, в RetinaFace контекстные модули, построенные на расширенных свёртках, помогают улучшить качество детекции.

В задачах детекции обычно используются комбинации функций потерь: для классификации (лицо/не лицо) – кросс-энтропия или Focal Loss (эффективна при дисбалансе классов), для регрессии координат рамок – Smooth L1 loss.

Выравнивание лиц с помощью CNN

После обнаружения лицо необходимо выровнять – привести к каноническому виду путем детектирования и нормализации по ключевым точкам. Эти точки, такие как углы глаз, кончик носа, углы рта, контуры бровей и подбородка, служат ориентирами для геометрической трансформации. Цель выравнивания – не просто найти точки, а использовать их для геометрической трансформации (например, аффинной или подобия) исходного изображения лица. Это приводит лицо к стандартному масштабу, ориентации и положению, где, например, глаза находятся на определенной горизонтальной линии, а расстояние между ними фиксировано. Такой канонический вид минимизирует вариации, не связанные с идентичностью, такие как поза и масштаб.

Выравнивание значительно снижает вариативность входных данных для последующей сети извлечения признаков, что напрямую повышает точность и робастность распознавания. Без выравнивания даже небольшие изменения в ракурсе или масштабе лица могут привести к значительным изменениям в векторе признаков, затрудняя сравнение.

Количество детектируемых ключевых точек варьируется в зависимости от задачи и необходимой точности:

– 5 точек: обычно это центры глаз, кончик носа и углы рта. Этого достаточно для базового выравнивания (например, в MTCNN или RetinaFace для последующей грубой нормализации).

– 68 точек: Стандарт, часто используемый в академических исследованиях (например, датасет iBUG 300-W), позволяет более точно смоделировать контур лица, форму глаз, бровей и губ.

– 98 точек и более: обеспечивают еще более детальное представление, включая контуры зрачков, век, более плотные точки на губах и контуре лица.

Большинство современных методов выравнивания основаны на регрессии координат ключевых точек с помощью CNN. Существуют два основных подхода [6]:

1. Прямая регрессия координат: Сеть напрямую предсказывает (x, y) координаты для каждой из N ключевых точек. В качестве функции потерь здесь обычно используются L1-норма (Mean Absolute Error) или L2-норма (Mean Squared Error), либо их комбинации, такие как Smooth L1 loss, которая менее чувствительна к выбросам, чем L2, и имеет более стабильные градиенты для малых ошибок, чем L1.

2.Регрессия тепловых карт: Сеть предсказывает N тепловых карт, по одной для каждой ключевой точки. Каждая тепловая карта представляет собой 2D распределение вероятностей, где пик яркости (высокое значение) указывает на наиболее вероятное положение соответствующей ключевой точки. Координаты точки затем извлекаются как

положение максимума на тепловой карте, часто с субпиксельной точностью (например, путем интерполяции или подгонки функции Гаусса). Этот подход часто более робастен и точен, особенно для сложных случаев, так как он неявно кодирует пространственную информацию и структуру вокруг точки и позволяет сети обучаться с более гладкими градиентами. Функции потерь для тепловых карт обычно основаны на MSE между предсказанной и эталонной (сгенерированной, например, 2D функцией Гаусса вокруг истинной координаты) тепловой картой. Продвинутые функции потерь, такие как Wing Loss и Adaptive Wing Loss, были предложены для улучшения точности, особенно для точек с большими отклонениями, уделяя больше внимания ошибкам на границах объектов и уменьшая чувствительность к малым ошибкам вблизи центра.

Архитектурно это могут быть как отдельные CNN, так и модули внутри более крупных сетей. Например, O-Net в MTCNN или специальная ветка в RetinaFace одновременно с детекцией предсказывают 5 ключевых точек. Это эффективно, так как признаки, извлеченные для детекции, частично полезны и для локализации точек. Для предсказания большого числа точек (68+) часто используются более глубокие и сложные архитектуры. Популярны архитектуры на основе модификаций ResNet или Hourglass Networks. Hourglass Networks, благодаря своей многомасштабной архитектуре с последовательными этапами понижения и повышения разрешения и пропусками соединений, эффективно агрегируют как глобальные контекстные, так и локальные высокодетализированные признаки, что критично для точной локализации всех точек. Другие архитектуры, такие как High-Resolution Networks (HRNet), стремятся поддерживать представление с высоким разрешением на протяжении всей сети, объединяя параллельные сверточные потоки разного разрешения, что также способствует высокой точности локализации.

В некоторых продвинутых системах выравнивание может также включать оценку 3D-положения головы (углы Эйлера: рыскание, тангаж, крен) по 2D-изображению с помощью CNN. Эти параметры затем могут использоваться для выполнения более сложной 3D-нормализации лица, проецируя его на фронтальный вид или используя их как дополнительные признаки.

Распознавание лиц: извлечение признаков, идентификация и верификация

Это ключевой этап, где выровненное изображение лица преобразуется в компактный и дискриминативный вектор признаков, который затем используется для идентификации или верификации личности.

Основная цель извлечения признаков – получить вектор, который будет дискриминативным (векторы признаков разных людей должны быть хорошо разделимы в пространстве признаков) и инвариантным (векторы признаков одного и того же человека должны быть близки, несмотря на изменения освещения, выражения лица, возраста и других неидентификационных факторов).

Для извлечения признаков часто используются следующие базовые архитектуры CNN:

– VGGNet: использовалась в ранних системах (например, VGG-16 (см. рис. 2)), но сейчас уступила место более глубоким архитектурам.

– ResNet (Residual Networks): является де-факто стандартом благодаря возможности обучать очень глубокие сети (ResNet-50 (см. рис. 3), ResNet-100 и глубже).

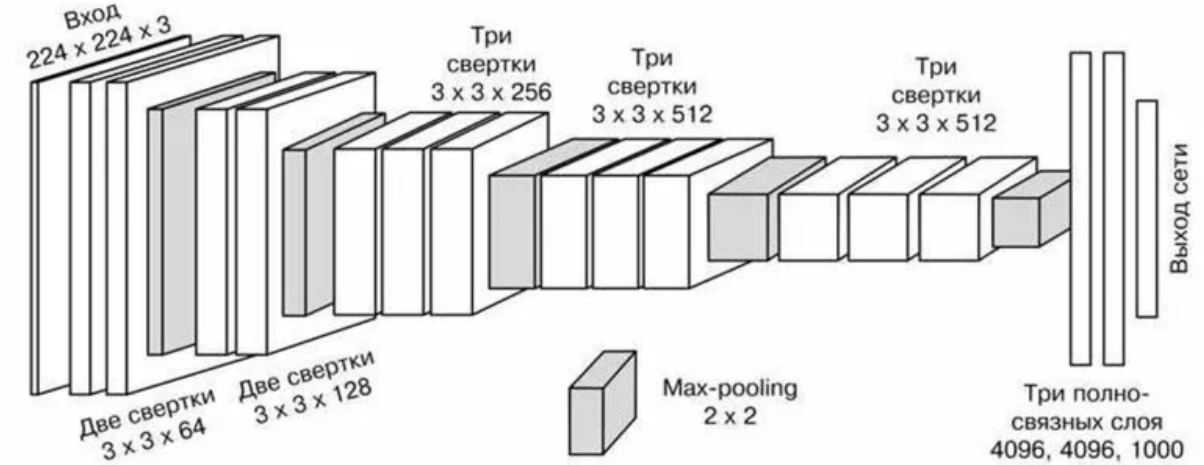

Рисунок 2 – Структура VGG-16

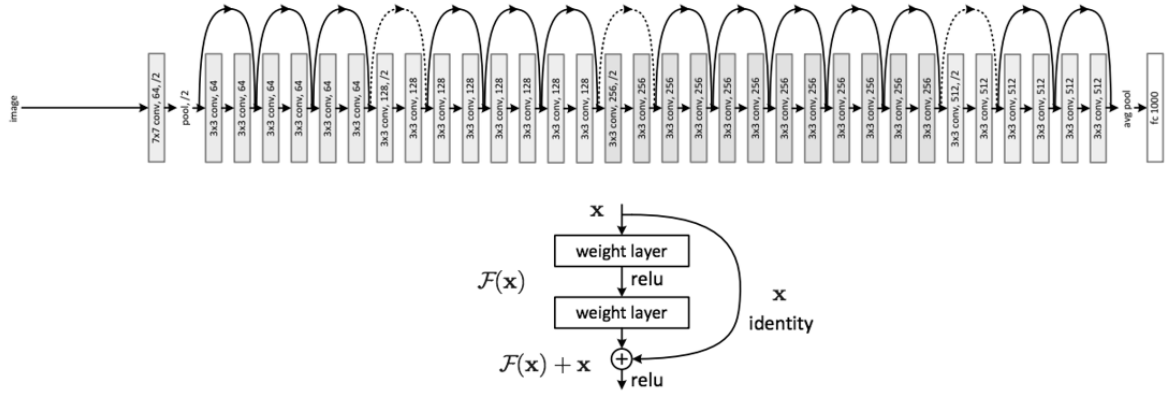

Рисунок 3 – Архитектура свёрточной нейронной сети ResNet50

– Inception: Архитектуры с инцепшн-модулями (например, в GoogLeNet (см. рис. 4)) эффективны благодаря параллельной обработке признаков на разных масштабах [7].

– Легковесные архитектуры: MobileNets, EfficientNets, ShuffleNets. Используют методы вроде глубинно-разделимых свёрток для снижения вычислительной сложности, что критично для мобильных устройств (например, MobileNet v2, представленная на рисунке 5) [8].

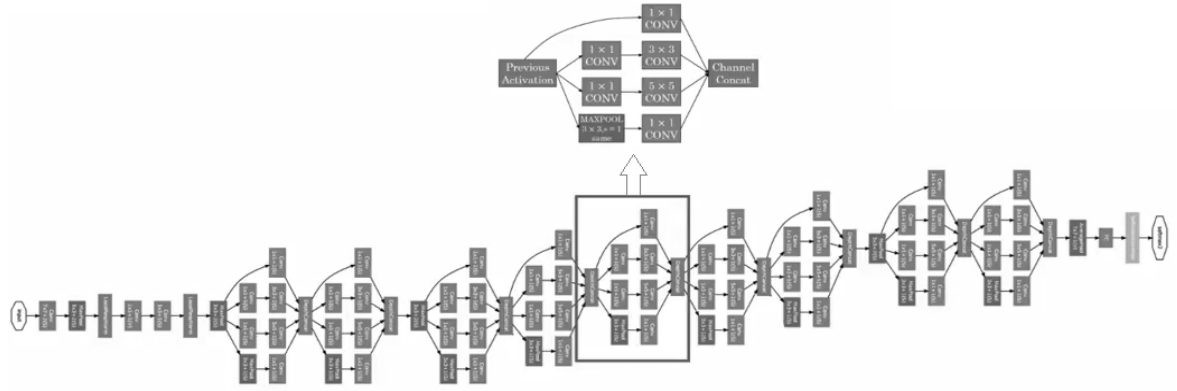

Рисунок 4 – Архитектура свёрточной нейронной сети Inception (GoogLeNet)

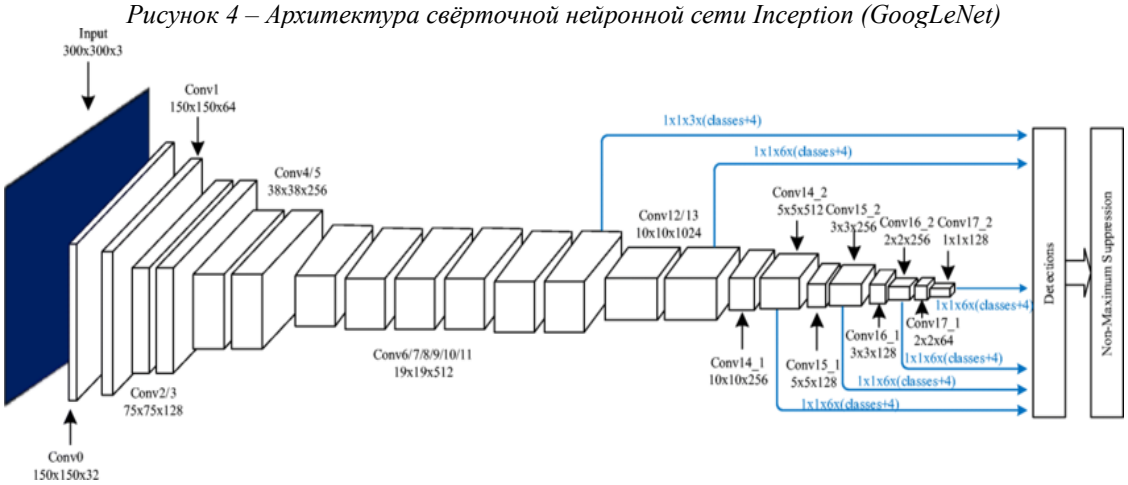

Рисунок 5 – Структура сети Mobilenet v2 + SSD

Для обучения дискриминативных признаков разработаны специальные функции потерь, работающие поверх выходов backbone-сети. К ним относятся Contrastive Loss и Triplet Loss, направленные на минимизацию расстояния между признаками одного класса и максимизацию расстояния между признаками разных классов. Также популярны модификации стандартной Softmax-потери с введением отступов (margins) для усиления внутриклассовой компактности и межклассовой разделимости, такие как SphereFace (A-Softmax), CosFace и ArcFace (Additive Angular Margin Loss), последняя является одной из наиболее популярных на сегодняшний день.

После получения векторов признаков следует этап принятия решения:

– Верификация: Задача сравнения "один к одному" (1:1). Вектор признаков предъявленного лица сравнивается с эталонным вектором признаков заявленной личности (например, хранящимся в базе данных). Если расстояние между векторами (например, евклидово или косинусное сходство) соответствует заданному порогу, личность подтверждается.

– Идентификация: Задача сравнения "один ко многим" (1:N). Вектор признаков предъявленного лица сравнивается со всеми векторами признаков в базе данных. Личность определяется как та, чей эталонный вектор наиболее близок к предъявленному (например, имеет наименьшее расстояние или наибольшее косинусное сходство), при условии, что это сходство превышает определенный порог.

Сами по себе эти этапы сравнения обычно не требуют специализированных CNN, а полагаются на метрики расстояния и пороговые значения. Однако качество и дискриминативность признаков, извлеченных CNN, напрямую определяют точность верификации и идентификации.

Для повышения робастности к окклюзиям, маскам и другим искажениям на этапе извлечения признаков или предобработки применяются различные подходы [9].

Аугментация данных включает применение различных преобразований к обучающим изображениям. Частичные свёртки могут использоваться в сетях для задач восстановления окклюдированных частей лица перед извлечением признаков или для обучения сетей быть более устойчивыми к отсутствующим данным; свёртка применяется только к валидным пикселям, а результат нормализуется.

Стробированные свёртки применяются в генеративных моделях и задачах реконструкции; обучаемый механизм стробирования может помочь сети адаптивно фокусироваться на значимых участках лица, игнорируя или восстанавливая искаженные, что косвенно улучшает качество извлекаемых признаков для распознавания.

Интегрированные и многозадачные CNN-архитектуры

Вместо последовательного применения отдельных моделей для каждого этапа, современные системы часто используют интегрированные архитектуры, выполняющие несколько задач одновременно.

Классическим примером интегрированной системы является MTCNN (Multi-task Cascaded Convolutional Networks) [4] (см. рис. 6).

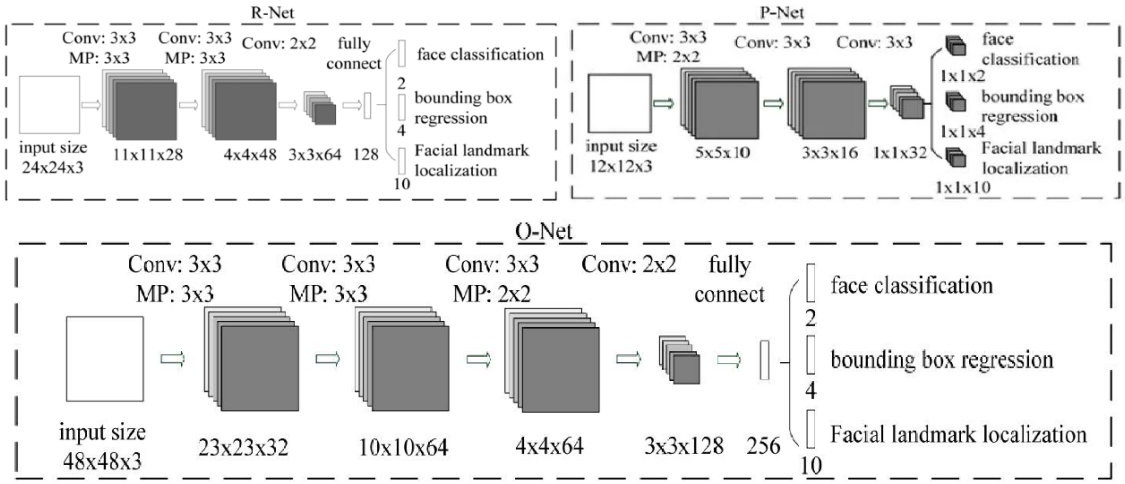

Рисунок 6 – Система MTCNN

Это каскад из трёх CNN: P-Net (Proposal Network) генерирует кандидатов окон с лицами; R-Net (Refine Network) фильтрует кандидатов и уточняет рамки; O-Net (Output Network) финально уточняет рамки и предсказывает координаты 5 ключевых точек лица для выравнивания. Каждая сеть решает несколько задач: классификация (лицо/не-лицо), регрессия ограничивающей рамки, регрессия ключевых точек (для O-Net).

Другим примером является RetinaFace (см. рис. 7) – современный одноэтапный детектор, который также является многозадачным [5]. На основе общего основной сети (например, ResNet+FPN) RetinaFace одновременно классифицирует области на наличие лица, регрессирует координаты ограничивающей рамки и координаты 5 ключевых точек лица. Иногда дополнительно предсказывает 3D-параметры лица или плотность пикселей лица. Преимуществами интегрированных подходов являются потенциально более высокая скорость (один проход через общую часть сети), лучшее использование признаков (признаки, полезные для одной задачи, могут быть полезны и для другой) и меньший размер общей модели.

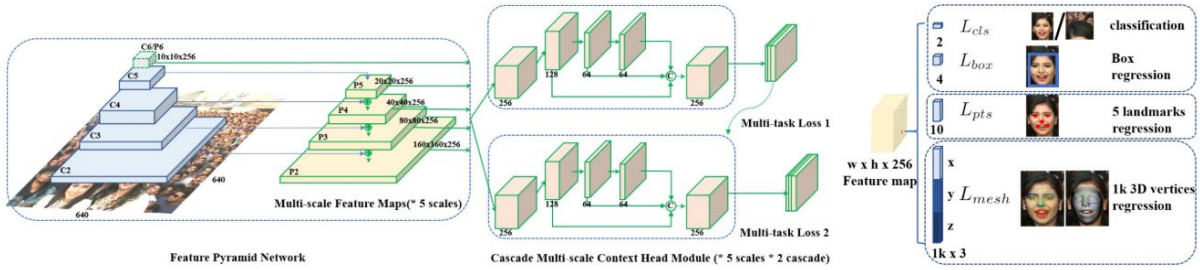

Рисунок 7 – Структура сети RetinaFace

Однако такие подходы сопряжены с вызовами, такими как сложность обучения и балансировки различных функций потерь, а также потенциальная интерференция между задачами. Хотя обнаружение и выравнивание часто интегрируются, этап извлечения признаков для распознавания (и последующей идентификации/верификации) обычно выполняется отдельной, более глубокой и специализированной CNN (например, ResNet с ArcFace) после того, как лицо было обнаружено и выровнено интегрированной системой.

Современные тенденции

Развитие CNN для анализа лиц продолжается, фокусируясь на следующих направлениях:

– Механизмы внимания: интеграция модулей, таких как Squeeze-and-Excitation (SE) блоки или Convolutional Block Attention Module (CBAM), позволяет сетям динамически перераспределять вычислительные ресурсы, фокусируясь на наиболее информативных признаках изображения и подавляя шум. Трансформерные модули самовнимания также начинают применяться для захвата глобальных зависимостей между частями лица, улучшая понимание контекста. Это особенно важно для распознавания лиц в сложных условиях (окклюзии, неоптимальные ракурсы).

– Трансформеры в зрении (Vision Transformers, ViT): показывают конкурентоспособные результаты, появляются гибридные модели (CNN+Transformer) типа ConvNeXt.

– 3D-распознавание и защита: Переход от 2D к 3D-анализу лиц позволяет повысить точность распознавания, особенно при вариациях позы, и является ключевым для борьбы со спуфинг-атаками (предъявление фотографии, видео или 3D-маски вместо реального лица). CNN обучаются извлекать признаки из 3D-данных (например, карт глубины, облаков точек) или оценивать 3D-форму лица по 2D-изображению. Для защиты от подделки разрабатываются CNN, анализирующие текстурные паттерны, микродвижения, отражательные свойства кожи и другие признаки "живости".

– Обучение на малых данных (Few-Shot Learning и One-Shot Learning) [10] представляет собой направление, ориентированное на разработку моделей, способных распознавать лица новых людей, имея лишь ограниченное число примеров. Суть подхода заключается не в запоминании конкретных классов, а в обучении самой способности быстро адаптироваться к новым задачам. Это достигается за счёт так называемого эпизодического обучения: модель во время тренировки решает множество небольших задач, имитирующих будущие тестовые условия, например, когда требуется различить пять лиц по одному примеру каждого. В качестве архитектурных решений часто применяются модели, основанные на сравнении эмбеддингов изображений. Например, в сиамских сетях модель учится определять, принадлежат ли два изображения одному и тому же человеку. В прототипических сетях каждое лицо представляется как вектор в признаковом пространстве, а классификация новых примеров выполняется по расстоянию до "центра" соответствующего класса. Также применяются методы метаобучения, такие как MAML (Model-Agnostic Meta-Learning), где модель обучается таким образом, чтобы её параметры можно было быстро адаптировать к новой задаче с минимальным числом градиентных шагов.

– Генеративные модели (GANs) находят применение для синтеза фотореалистичных изображений лиц, что полезно для аугментации данных, особенно для редких ракурсов, выражений или демографических групп. Они также используются для задач нормализации изображений (например, изменение ракурса лица к фронтальному, удаление очков, омоложение/состаривание), что может улучшить каноничность входных данных для основной сети распознавания.

Выводы

Свёрточные нейронные сети играют центральную роль на каждом этапе систем распознавания лиц. Для обнаружения и выравнивания часто используются

специализированные или многозадачные CNN, такие как MTCNN и RetinaFace, применяющие в том числе расширенные свёртки для лучшего анализа контекста. Для извлечения дискриминативных признаков, используемых в дальнейшем для идентификации и верификации, применяются глубокие архитектуры типа ResNet в сочетании с продвинутыми функциями потерь (ArcFace, CosFace). Облегчённые модели (MobileNet) обеспечивают работу на мобильных устройствах. Использование частичных и стробированных свёрток помогает в обработке неидеальных данных и генерации. Будущее, вероятно, за дальнейшей интеграцией, гибридными моделями и решением проблем робастности и скорости работы.

Литература

1. Федяев, О. И. Автоматическая регистрация присутствия студентов на учебном занятии с помощью компьютерного зрения / О. И. Федяев, И. А. Коломойцева // XXI Национальная конференция по искусственному интеллекту с международным участием КИИ-2023 (Смоленск, 16-20 октября 2023 г.). Труды конференции. В 2-х томах. Т.1. – Смоленск: Принт-Экспресс, 2023. – С. 294-303.

2. Du L., Zhang R., Wang X. Overview of two-stage object detection algorithms //Journal of Physics: Conference Series. – IOP Publishing, 2020. – Т. 1544. – №. 1. – С. 012033.

3. Zhang Y. et al. A comprehensive review of one-stage networks for object detection //2021 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC). – IEEE, 2021. – С. 1-6.

4. Zhang N., Luo J., Gao W. Research on face detection technology based on MTCNN //2020 international conference on computer network, electronic and automation (ICCNEA). – IEEE, 2020. – С. 154-158.

5. Deng J. et al. Retinaface: Single-shot multi-level face localisation in the wild //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. – 2020. – С. 5203-5212.

6. Wang X., Bo L., Fuxin L. Adaptive wing loss for robust face alignment via heatmap regression //Proceedings of the IEEE/CVF international conference on computer vision. – 2019. – С. 6971-6981.

7. Sam S. M. et al. Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3 //Procedia Computer Science. – 2019. – Т. 161. – С. 475-483.

8. Zhang X. et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2018. – С. 6848-6856.

9. Alzubaidi L. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions //Journal of big Data. – 2021. – Т. 8. – С. 1-74.

10. Holkar A., Walambe R., Kotecha K. Few-shot learning for face recognition in the presence of image discrepancies for limited multi-class datasets //Image and Vision Computing. – 2022. – Т. 120. – С. 104420.

Кочетуров В. В., Федяев О. И. Свёрточные нейронные сети в системах обнаружения и распознавания лиц. *В статье рассматривается применение свёрточных нейронных сетей на различных этапах систем распознавания лиц: обнаружение, выравнивание и распознавание. Обозреваются ключевые архитектуры и методы, применяемые на каждом из этапов. Описываются многозадачные CNN, выполняющие несколько функций одновременно. Выделяются современные тенденции развития: применение механизмов внимания, трансформеров и обучение на малых данных.*

Ключевые слова: *свёрточные нейронные сети, распознавание лиц, обнаружение лиц, выравнивание лиц, архитектуры сетей, компьютерное зрение.*

Kocheturov V.V., Fedyaev O.I. Convolutional neural networks in face detection and recognition systems. *The article discusses the application of convolutional neural networks at different stages of face recognition systems: detection, alignment and recognition. The key architectures and methods used in each stage are reviewed. Multitasking CNNs that perform multiple functions simultaneously are described. Current development trends are highlighted: the application of attention mechanisms, transformers, and learning from small data.*

Keywords: *convolutional neural networks, face recognition, face detection, face alignment, network architectures, computer vision.*

Статья поступила в редакцию 28.02.2025
Рекомендована к публикации профессором Зори С. А.