Обнаружение объектов на изображении: от критериев Байеса и Неймана–Пирсона к детекторам на базе нейронных сетей EfficientDet

Н.А. Андриянов [1], В.Е. Дементьев [2], А.Г. Ташлинский [2]
[1] Финансовый университет при Правительстве Российской Федерации, 125993, Россия, г. Москва, Ленинградский пр-т, д. 49;
[2] Ульяновский государственный технический университет, 432027, Россия, г. Ульяновск, ул. Северный Венец, д. 32

Аннотация

Актуальность задач обнаружения и распознавания объектов на изображениях и их последовательностях с годами только возрастает. За последние несколько десятилетий предложено огромное количество подходов и методов обнаружения как аномалий, то есть областей изображения, характеристики которых отличаются от прогнозных, так и объектов интереса, о свойствах которых есть априорная информация, вплоть до библиотеки эталонов. В работе предпринята попытка системного анализа тенденций развития подходов и методов обнаружения, причин этого развития, а также метрик, предназначенных для оценки качества и достоверности обнаружения объектов. Рассмотрено обнаружение на основе математических моделей изображений. При этом особое внимание уделено подходам на основе моделей случайных полей и отношения правдоподобия. Проанализировано развитие сверточных нейронный сетей, направленных на задачи распознавания и обнаружения, включая ряд предобученных архитектур, обеспечивающих высокую эффективность при решении данной задачи. В них для обучения используются уже не математические модели, а библиотеки реальных снимков. Среди характеристик оценки качества обнаружения рассмотрены вероятности ошибок первого и второго рода, точность и полнота обнаружения, пересечение по объединению, интерполированная средняя точность. Также представлены типовые тесты, которые применяются для сравнения различных нейросетевых алгоритмов.

Ключевые слова: распознавание образов, обнаружение объектов, компьютерное зрение, обработка изображений, случайные поля, CNN, IoU, mAP, вероятность правильного обнаружения.

Цитирование: **Андриянов, Н.А.** Обнаружение объектов на изображении: от критериев Байеса и Неймана–Пирсона к детекторам на базе нейронных сетей EfficientDet / Н.А. Андриянов, В.Е. Дементьев, А.Г. Ташлинский // Компьютерная оптика. – 2022. – Т. 46, № 1. – С. 139-159. – DOI: 10.18287/2412-6179-CO-922.

Citation: Andriyanov NA, Dementiev VE, Tashlinskii AG. Detection of objects in the images: from likelihood relationships towards scalable and efficient neural networks. Computer Optics 2022; 46(1): 139-159. DOI: 10.18287/2412-6179-CO-922.

Введение

Задача обнаружения объектов интереса и аномалий на видеопоследовательностях и отдельных изображениях остается актуальной уже не одно десятилетие, и число приложений, где она возникает, только возрастает. Среди них приложения, связанные с дистанционным зондированием Земли [1, 2], наземным мониторингом [3, 4], медицинскими исследованиями [5, 6], радиолокацией [7, 8], сельским хозяйством [9, 10], рентгеновским сканированием багажа [11, 12] и многие другие. Для обнаружения объектов интереса используется различная априорная информация вплоть до заданной библиотеки эталонов, а под аномалией (злокачественные образования, появление артефактов на видеопоследовательности и пр.), как правило, понимается некоторая область изображения, характеристики которой отличаются от прогнозных, сформированных в процессе обработки изображений. Чем больше размерность изображения, т.е. количество координат, которым описывается каждый пиксель (например, пространственные координаты, номер кадра в видеопоследовательности, номер спектрального диапазона), тем выше сложность решения этих задач.

При статистическом анализе изображений [13] их стохастической моделью обычно служат случайные поля, а задача обнаружения рассматривается в условиях шумов. В качестве предобработки наиболее часто применяется процедура фильтрации [14, 15]. Тогда при наличии априорной информации об обнаруживаемом сигнале и заданной математической модели фона можно воспользоваться оптимальным байесовским обнаружителем [16], минимизирующим средний риск, связанный с вероятностями ошибок первого и второго рода, называемых в ряде приложений вероятностями пропуска цели и ложной тревоги. Пространство наблюдений разделяется на две области: каждый анализируемый пиксель изображения приписывается либо области G_0, соответствующей состоянию S_0 отсутствия сигнала, либо области G_1,

соответствующей состоянию S_1 его наличия. Результатом обработки наблюдений является некая статистика L, а решающее правило принятия или отклонения гипотезы о наличии H_1 (или отсутствии H_0) сигнала сводится к сравнению численной характеристики данной статистики с пороговым значением. При этом дополнительно возникает задача выбора порога.

Для практических задач характерна разная степень априорной неопределенности, и, как правило, минимизируется одна из ошибок, а для другой задается некоторое граничное значение. Например, критерий Неймана–Пирсона [17] модифицирует байесовское правило обнаружения. В отсутствие априорной вероятности об ошибках данный критерий минимизирует вероятность ошибки первого рода при заданной вероятности ошибки второго рода. Подобные обнаружители хорошо себя зарекомендовали при обработке многомерных и многозональных изображений [18, 19], когда удается задать адекватную математическую модель, описывающую подстилающую поверхность (фон). Более того, такой подход также дает возможность обнаружения недетерминированных аномалий на основе модифицированного отношения правдоподобия [20]. С другой стороны, требуемые вероятности ошибок достигаются при адекватности используемых математических моделей изображений. При этом отсутствует некая универсальная модель, подходящая для описания изображений при сложной помеховой обстановке или большом числе аномалий.

Для ситуации, когда возможные типы объектов известны, новое качество решения задачи их обнаружения было достигнуто с появлением глубокого обучения и сверточных нейронных сетей CNN (сокращение от «Convolutional Neural Network») [21, 22]. Нейронная сеть производит настройки весов нейронов при обработке имеющихся наблюдений и выборе порога принятия решения. Использование сверточных сетей позволяет находить объекты разных размеров и классов. При этом возможны как минимум два варианта их применения. Во-первых, в ряде задач могут быть успешно использованы уже готовые или предобученные архитектуры CNN, которые способны распознавать только те объекты, на которых они обучались. Во-вторых, для многих более частных задач можно использовать новую базу изображений и объектов для того, чтобы дообучить модель с предобученной архитектурой (трансферное обучение) под конкретную узкую задачу [23]. Во втором случае настройка весов, как правило, происходит только для последних полносвязных слоев, все остальные коэффициенты модели не изменяются, поскольку служат в сети для извлечения признаков с картинки.

Развитие нейросетевых детекторов для задач обнаружения привело к появлению множества архитектур, в частности, глубоких нейронных сетей DNN (сокращение от «Deep Neural Network») [24]. Основным вектором их развития стало эффективное обнаружение в режиме реального времени. Особо можно выделить сеть R-CNN (от «Region Based»), направленную на распознавание объектов не на всем изображении, а в локальных областях изображений [25]. Архитектура R-CNN породила серию предметно ориентированных улучшенных моделей, в том числе: сеть Fast R-CNN (быстрая R-CNN) [26] для задачи классификации и регрессии покрывающего объект прямоугольника; сеть Faster R-CNN (ускоренная Fast R-CNN) [27], использующую вспомогательную подсеть для генерации регионов интереса.

Еще большую производительность демонстрируют сети архитектуры YOLO (сокращение от «You Only Look Once» – «ты смотришь только раз») [28], за один проход формирующие и ограничивающие регионы локализации объекта, и метку класса объект. Высокие быстродействие и производительность обеспечивает также сеть SSD (сокращение от «Single Shot Multibox Detector» –«один снимок – множество обнаружений») [29, 30], в основе которой дискретизация выходного пространства прямоугольных областей обнаружения в прямоугольники из стандартного набора с заданными размерами для каждого местоположения на карте признаков (характерных особенностей) изображения.

В статье рассмотрены критерии качества алгоритмов обнаружения, классические статистические обнаружители, использующие математические модели случайных полей, а также современные решения на базе архитектур CNN из популярного фреймворка для машинного обучения TensorFlow [31], такие как R-CNN, FastR-CNN, FasterR-CNN, YOLO, EfficientDet, SSD и др. Приведен взгляд авторов на основные перспективы и тенденции в задаче обнаружения объектов на изображениях.

В целом, анализ доступной литературы показывает, что в настоящее время на русском языке отсутствуют обзорные работы, которые бы достаточно полно охватывали данную тематику. У зарубежных ученых имеются интересные обзорные публикации, но данная область развивается достаточно быстро, что приводит к устареванию некоторых алгоритмов. Например, в работе [32] описываются статистические методы распознавания образов, но не уделяется должного внимания алгоритмам на базе глубокого обучения. В работе [33] авторы уделяют основное внимание методам на базе глубокого обучения и не рассматривают альтернативные подходы. Однако данный обзор требует от читателя достаточно глубоких предварительных знаний по обнаружителям, например, метрик, используемых для оценки качества моделей. В обзорной публикации [34] выделены общие для всех систем обнаружения компоненты, такие как наличие базы моделей, детектор признаков, гипотезы об объектах и верификатор гипотез. Несмотря на то, что обзор подходит для начинающих исследователей, многие алгоритмы в нем рассматриваются доста-

точно поверхностно, а сами рассматриваемые методы и модели не всегда показывают наилучшие результаты на практике.

Актуальность тематики также подтверждается ростом числа статей в области компьютерного зрения. Согласно [35], за период с 1998 по 2018 гг. количество публикаций по тематике обнаружения объектов на изображениях увеличилось в 60 раз, и тенденция к увеличению также сохраняется. На рис. 1 показана диаграмма из соответствующего источника.

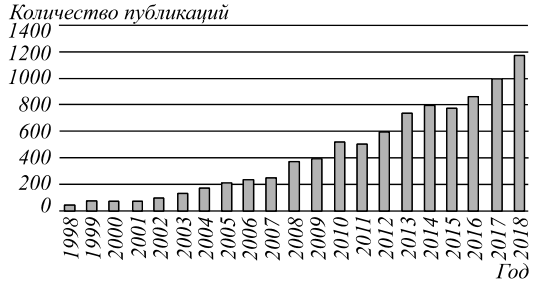

Рис. 1. Анализ публикационной активности по тематике [35]

Более того, по данным Gartner [36], «размер мирового рынка компьютерного зрения оценивался в 10,6 млрд. долларов США в 2019 году и, как ожидается, будет расти со среднегодовым темпом роста (CAGR) в 7,6 % с 2020 по 2027 год». Причиной этого являются острая потребность в современной автоматизации в обрабатывающей промышленности; резкий рост спроса на системы визуального контроля качества. Эксперты ожидают экспоненциального роста рынка, как показано на рис. 2 [36]. При этом одно из ведущих направлений систем компьютерного зрения – это именно интеллектуальные системы обнаружения и распознавания образов.

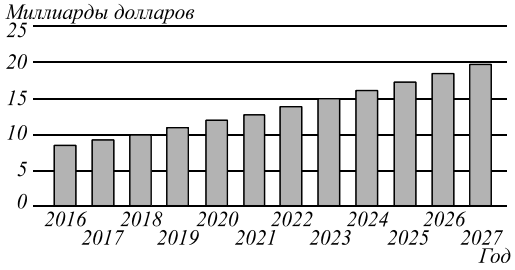

Рис. 2. Оценка рынка компьютерного зрения [36]

Таким образом, тематика обнаружения и распознавания объектов в настоящее время является актуальной. Данная статья, с одной стороны, описывает основные характеристики и метрики, используемые в обнаружителях, рассматривает алгоритмы обнаружения аномалий и, с другой стороны, подробно описывает методы глубокого обучения, используемые уже в задаче обнаружения известных объектов.

1. Критерии эффективности обнаружения объектов

Поскольку объекты на изображениях могут быть разных типов, задача их обнаружения часто связана с задачей распознавания. Обнаружение можно свести к различению гипотез о наличии объекта H_1 и его отсутствии H_0. Причем, как правило, гипотезы выдвигаются относительно не всего изображения, а некоторой его локальной области [37 – 39]. Тогда принять решение об отсутствии объекта на всем изображении можно, лишь проверив все возможные области на этом изображении.

Как уже отмечалось ранее, проверка гипотез H_1 и H_0 может привести к ошибкам первого (правильная гипотеза отклоняется) и второго (принимается неверная гипотеза) рода. На практике обнаружитель настраивается под конкретные требования, связанные с критичностью той или иной ошибки. Например, при автоматизированном досмотре минимизируется число ошибок первого рода, а при определении заканчивающегося на полках магазина товара – число ошибок второго рода. При этом во многих прикладных задачах решение требуется принимать в режиме, близком к реальному времени.

Одним из вариантов решения задачи обнаружения является полный перебор всех областей изображения. При этом требуется минимизировать вероятность ошибки одного рода при заданном пороговом значении вероятности ошибки другого рода. Критериями эффективности такого подхода, в первую очередь, выступают доля правильных обнаружений и производительность алгоритма обнаружения. Последняя характеризуется числом стандартных кадров, обрабатываемых в единицу времени, либо временем, затрачиваемым на обработку стандартного кадра. Это приводит к задачам предметно ориентированной оптимизации алгоритмов обнаружения [40].

Кроме указанных критериев, алгоритмы обнаружения характеризуются корректным определением самой области, содержащей объект. В частности, метрикой *IoU* (сокращение от «Intersection over Union» – «пересечение по объединению») [41]. Вообще говоря, значение *IoU* представляет собой отношение площади, полученной в результате пересечения области, предсказанной алгоритмом, и реальной области с объектом (аномалией), к площади, полученной в результате объединения этих областей. Геометрический смысл метрики иллюстрирует рис. 3, из которого видно, что если области совпадают, то значение $IoU=1$, а если не пересекаются, то $IoU=0$. На рисунке $\overline{j}_{А1}$ и $\overline{j}_{А2}$ – координаты левого верхнего и правого нижнего углов спрогнозированной области объекта (обозначенной на рисунке «А»), $\overline{j}_{Б1}$ и $\overline{j}_{Б2}$ – аналогичные координаты области истинного положения объекта (обозначенной на рисунке «Б»). Понятно, что этот критерий требует знания положения на изображении действительной области с объектом (аномалией), которая, например, может быть отмечена экспертом.

В настоящее время в компьютерном зрении все большее применение находит метрика *mAP* (сокращение от «Mean Average Precision» – интерполированная

средняя точность) [42], учитывающая, кроме точности прогноза областей объектов (аномалий), еще и достоверность правильного обнаружения, что позволяет сравнивать эффективность различных алгоритмов обнаружения. Достоверность правильного обнаружения определяется точностью (precision), характеризующей способность алгоритма обнаруживать именно нужные объекты (аномалии), и полнотой (recall) обнаружения [43], характеризующей способность алгоритма находить нужный объект (аномалию) в полном объеме (в идеале все объекты заданного класса):

$$precision = \frac{TP}{TP + FP}, \quad (1)$$

$$recall = \frac{TP}{TP + FN}, \quad (2)$$

где TP – число правильных решений о наличии объекта (аномалии); FP – число ошибок второго рода; FN – число ошибок первого рода.

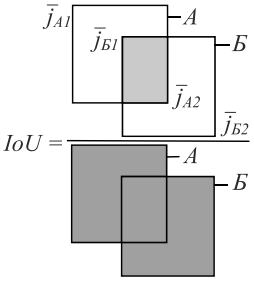

Рис. 3. Пояснение смысла метрики IoU

Геометрический смысл понятий полноты и точности поясняет рис. 4, где TN – число правильных решений об отсутствии объекта (аномалии).

Отметим, что совокупность двух характеристик (точности и полноты) дает более глубокое понимание качества работы алгоритма обнаружения по сравнению с расчетом доли верных обнаружений (accuracy):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (3)$$

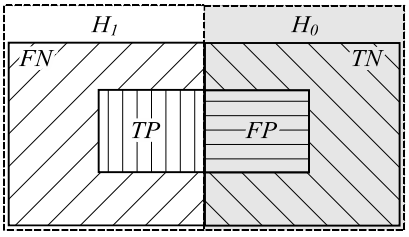

Рис. 4. Геометрическая интерпретация понятий полноты и точности

Особенно ярко это проявляется для несбалансированных данных, где количество объектов одного класса может значительно превышать количество объектов другого класса, и показатель (3) не позволяет адекватно оценить качество алгоритма.

Для интегральной характеристики точности и полноты используют также F-меру [44]. Данная метрика рассчитывается в соответствии с выражением (4).

$$F_\beta = (1+\beta^2)\frac{precision \times recall}{\beta^2 \times precision + recall}, \quad (4)$$

где β – весовой коэффициент, определяющий вклад точности в значение F-меры.

Для расчета *mAP* используется метод, аналогичный построению ROC-кривых (сокращение от «Receiver Operating Characteristic Curve» – «кривая рабочих характеристик», используемая для оценки качества бинарной классификации) [45]. Однако здесь требуется построить зависимость точности от полноты, так называемую PR-кривую (сокращение от «Precision-Recall Curve» – «кривая точности-полноты») [46]. Обычно для полноты (ось абсцисс) выбирают некоторый шаг в диапазоне от 0 до 1. Например, в соревновании Pascal VOC 2008 для оценки mAP рассчитывается интерполированная PR-кривая из 11 точек с шагом по полноте 0,1. Пример такой кривой на основе данных из [47] показан на рис. 5. Из рисунка видно, что на интервалах монотонного поведения PR-кривой допускается аппроксимация максимальным значением (пунктирная линия). Обычно это делается для уменьшения зависимости *mAP* от вариаций кривой. Значение *mAP* соответствует площади под PR-кривой (от 0 до 1). Поэтому аппроксимация может привести к некоторому завышению средней точности. Отметим, что *mAP* рассчитывается при условии, что параметр *IoU* превышает некоторый заданный порог λ. В частности, при расчете приведенной кривой $\lambda = 0{,}5$.

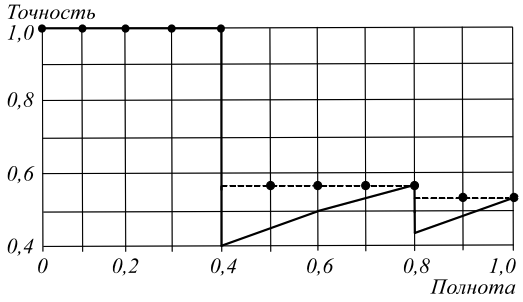

Рис. 5. Пример PR-кривой

Для оценки качества алгоритмов обнаружения используются и другие, близкие по смыслу метрики. В частности, в [48] предложено 12 метрик для датасета COCO (сокращение от «Common Objects in Context» – «реальные объекты в контексте»), содержащего размеченные изображения различных положений объектов на различных фонах. Кроме средней точности, может быть использована также средняя полнота, причем расчет этих параметров проводится с учетом трех возможных диапазонов размера объекта и трех ограничений на количество объектов на изображении.

Если требуется обнаружить только один объект, то качество его обнаружения можно оценить объектовой средней точностью *AP* [49]. Можно выделить три подхода к ее расчету.

1) Задается вектор пороговых значений $\bar{\lambda}$ для параметра *IoU*. Для каждого порога λ_i находится число TP_{IoU} правильных обнаружений, соответ-

ствующее условию $IoU \geq \lambda_i$, а также число ошибок первого и второго рода. Затем по выражениям (1) и (2) рассчитываются вектора точности и полноты. Затем компоненты вектора полноты упорядочиваются по возрастанию (ось абсцисс) и им в соответствие ставятся значения точности (ось ординат), строится интерполированная PR-кривая, и определяется *AP* как площадь под ней.

2) Для фиксированного значения $\lambda = 0{,}5$ порога параметра *IoU* выбираются разные пороги обнаружения. Их совокупность называют вектором $\overline{C}$ уверенностей обнаружителя. Например, одна и та же область на изображении может быть похожа на объект первого типа на X% и на объект второго типа на Y%. Если $X/100 \geq C_i$, то принимается решение об обнаружении объекта первого типа при i-м пороге. Аналогично и для $Y/100 \geq C_i$. Если X и Y одновременно превышают заданный порог, то метка класса ставится в соответствии с максимальным значением из X и Y. Далее, как и для первого подхода, рассчитываются количественные характеристики обнаружения и по соотношениям (1) и (2) получают вектора значений точности и полноты, строят PR-кривую и определяют объектовую точность *AP*.

3) Аналогично второму подходу задается вектор уверенностей $\overline{C}$. Затем для каждого объекта и каждого порога C_i рассчитываются качественные характеристики обнаружения. При этом объектовая точность определяется, как отношение площади, полученной в результате пересечения ограничивающих прямоугольников реального расположения объекта и прогнозируемого, к площади спрогнозированного ограничивающего прямоугольника, а полнота – к площади ограничивающего прямоугольника реального расположения объекта (рис. 6, где смысл областей А и Б аналогичен рис. 3). По всем объектам находятся средние значения характеристик, являющиеся компонентами векторов, строится PR-кривая, по которой определяют *AP*.

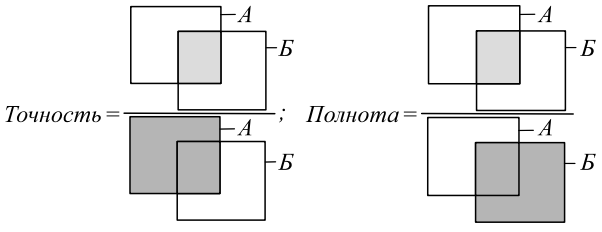

Рис. 6. Пояснение к понятиям точности и полноты

Если указанные действия выполнены для одного класса объектов, то получаем среднюю точность *AP* обнаружения объектов данного класса, если для множества классов, то получаем *mAP*, как среднее арифметическое объектовых средних точностей каждого класса.

Следует отметить, что при описанном подходе число классов считается известным, а размеры объектов и их расположение – некоррелированными.

Интересный подход на базе фильтров-локализаторов рассмотрен в трудах Л.П. Ярославского [50]. Для локализации объектов на изображениях автором было предложено использовать адаптивные линейные фильтры. При этом такой подход на сложных изображениях позволил добиться лучших результатов, чем у стандартных обнаружителей, базирующихся на расчете корреляции. В работе [50] также приводятся два типа ошибок для оценки качества алгоритмов измерения координат объекта на изображении:

1) Ошибки первого рода [50] связаны с тем, что отдельные детали изображения неверно принимаются за объекты, которые необходимо обнаружить на изображении. Как правило, для ошибок данного рода свойственно большое расхождение оцененных координат с истинными, поэтому они получили название аномальных ошибок.

2) Ошибки второго рода [50] обычно не превышают по порядку размеры объекта, чаще всего связаны с повреждением сигнала объекта обнаружения некоторым шумом, например, шумом регистратора изображения. Ошибки данного рода были названы нормальными.

Резюмируя первый параграф, отметим, что существует большое количество метрик, которые используются для оценки качества алгоритмов обнаружения, поэтому их применение должно обосновываться целесообразностью и подбираться индивидуально в конкретных задачах.

2. Обнаружение на основе математических моделей изображений

При обнаружении сигналов на базе математической модели чаще всего используется модель наблюдений в виде аддитивной смеси коррелированного сигнала (фона) и белого шума [51 – 54]:

$$z_{\overline{j}} = x_{\overline{j}} + \theta_{\overline{j}}, \overline{j} \in J, \tag{5}$$

где $\{x_{\overline{j}}\}$ – случайное поле, описывающее коррелированную яркость подстилающей поверхности; $\theta_{\overline{j}}$ – белый шум; $\overline{j} = (j_1, j_2, ..., j_N)$ – вектор координат сетки отсчетов в N-мерном пространстве; $J - N$-мерная область, ограничивающая многомерное изображение, в частности для плоских изображений – прямоугольник.

В случае наличия в какой-то области изображения полезного детерминированного сигнала $s_{\overline{j}}$ модель (5) для нее изменится и примет вид (6).

$$z_{\overline{j}} = \begin{cases} x_{\overline{j}} + \theta_{\overline{j}}, \overline{j} \notin G_1, \\ x_{\overline{j}} + s_{\overline{j}} + \theta_{\overline{j}}, \overline{j} \in G_1, \end{cases} \tag{6}$$

где G_1 – область сигнала $s_{\overline{j}}$.

Алгоритм обнаружения, как правило, подразумевает реализацию двух этапов [55]: первый – декорреляция составляющей фона, второй – обнаружение

сигнала на декоррелированной подстилающей поверхности.

Декорреляция фона обычно реализуется с помощью построения прогноза яркости $\hat{x}_{\bar{j}}$ подстилающей поверхности для каждого узла $\bar{j}$ сетки отсчетов и его вычитания из наблюдений $z_{\bar{j}}$. Прогноз выполняется по имеющимся наблюдениям и предполагаемой модели фона. При этом используется линейное взвешенное суммирование наблюдений в некоторой (как правило, небольшой) окрестности M точки $\bar{j}_0$ прогнозируемой яркости [56], определяемое выражением (7).

$$\hat{x}_{\bar{j}_0} = \sum_{\bar{k}\in M} \alpha_{\bar{k}} z_{\bar{j}_0-\bar{k}}\,, \tag{7}$$

где M – область, определяющая окрестность формирования прогноза; $\alpha_{\bar{k}}$ – весовые коэффициенты.

Таким образом, для построения оптимального по некоторому критерию прогноза требуется найти «оптимальные» коэффициенты, например, в смысле минимума дисперсии ошибки прогноза.

Можно выделить несколько критериев обнаружения сигналов на основе статистических моделей.

Самое широкое распространение получили критерии Байеса и Неймана–Пирсона. В них для принятия решения о наличии или отсутствии сигнала находят отношение условных плотностей распределения вероятностей (ПРВ) наблюдений, также называемое отношением правдоподобия (8):

$$L = \frac{\omega\left(\{z_{\bar{j}}\}\,|\,H_1\right)}{\omega\left(\{z_{\bar{j}}\}\,|\,H_0\right)}, \tag{8}$$

где $\omega(\{z_{\bar{j}}\}\,|\,H_1)$ и $\omega(\{z_{\bar{j}}\}\,|\,H_0)$ – условные ПРВ наблюдений при наличии и отсутствии сигнала (в условиях справедливости гипотезы H_1 и H_0 соответственно).

Затем отношение правдоподобия L сравнивают с пороговым значением λ_0, реализуя решающее правило [19] в виде (9):

$$L\begin{cases} > \lambda_0, & \text{сигнал есть},\\ \le \lambda_0, & \text{сигнала нет}. \end{cases} \tag{9}$$

Недостатком байесовского решающего правила является необходимость наличия априорной информации о вероятностях состояния объекта. Однако даже если известны вероятности ошибок первого и второго рода обнаружителя, этого недостаточно для определения $\omega(\{z_{\bar{j}}\}\,|\,H_1)$ и $\omega(\{z_{\bar{j}}\}\,|\,H_0)$. Поэтому в подобных ситуациях используется критерий Неймана–Пирсона, в котором пороговое значение находится из условия, что при выбранном допустимом максимальном уровне вероятности ошибки ложной тревоги, решающее правило должно минимизировать вероятность ошибки пропуска цели [17, 18], а оценки ПРВ $\omega(\{z_{\bar{j}}\}\,|\,H_1)$ и $\omega(\{z_{\bar{j}}\}\,|\,H_0)$ получают с использованием параметрических и непараметрических методов [57]. Последние, т.е. непараметрические методы, чаще всего применяются при негауссовости условных ПРВ. Параметрический подход предполагает аппроксимацию указанных распределений Гауссовым законом. Это, с одной стороны, упрощает задачу, потому что достаточно оценить лишь два параметра: математическое ожидание и ковариацию [37]. С другой стороны, предположение о нормальности условных распределений накладывает ряд ограничений использования. Тем не менее, такой подход для многих прикладных задач показал свою эффективность.

В качестве параметрической модели ПРВ можно, например, использовать выражение (10).

$$\begin{aligned} &\omega\left(\{z_{\bar{j}}\}\,|\,H_{0,1}\right) \cong \frac{1}{\sqrt{2\pi\det(V)}}\times\\ &\times\exp\left\{-\frac{1}{2}\left(z_{\bar{j}}-m_{0,1\bar{j}}\right)V^{-1}_{\bar{j}\bar{k}}\left(z_{\bar{j}}-m_{0,1\bar{j}}\right)^T\right\}, \end{aligned} \tag{10}$$

где $m_{0\bar{j}} = M\{z_{\bar{j}}\,|\,H_0\} = \hat{x}_{\bar{j}} = M\{x_{\bar{j}}\,|\,Z_0\}$ – оптимальный в смысле минимума дисперсии ошибки прогноз случайного поля на основе всех наблюдений Z_0 из области G_0, в которой полезный сигнал заведомо отсутствует; $m_{1\bar{j}} = M\{z_{\bar{j}}\,|\,H_1, s\} = s_{\bar{j}} + \hat{x}_{\bar{j}}$ – оптимальный прогноз смеси коррелированного фона с сигналом; $V_{\bar{j}\bar{k}} = M\{(z_{\bar{j}} - m_{0,1\bar{j}})(z_{\bar{k}} - m_{0,1\bar{k}})^T\} = P_{\bar{j}\bar{k}} + \sigma_0^2 E_{\bar{j}\bar{k}}$ – ковариационная матрица наблюдений; $P_{\bar{j}\bar{k}} = M\{(x_{\bar{j}} - \hat{x}_{\bar{j}})(x_{\bar{k}} - \hat{x}_{\bar{k}})^T\}$ – ковариационная матрица ошибок при оптимальном прогнозировании; σ_0^2 – дисперсия шума; $E_{\bar{j}\bar{k}}$ – единичная матрица; $s_{\bar{j}}$ – значение полезного сигнала в точке с координатами $\bar{j}$.

На основе формулы (10) и правила (9) отношение правдоподобия может быть представлено в виде (11) [58]:

$$L = \sum_{\bar{j}}\sum_{\bar{k}\in G_0} s_{\bar{j}} V^{-1}_{\bar{j}\bar{k}}\left(z_{\bar{k}} - \hat{x}_{\bar{k}}\right). \tag{11}$$

Отметим, что при увеличении области сигнала G_1 задача построения прогноза $\hat{x}_{\bar{k}}$ усложняется, поскольку для каждой точки области, как правило, используется отдельная процедура вычисления. В работе [18] предложено решение, существенно упрощающее расчет отношения правдоподобия, в котором, в отличие от (11), используется прогноз не в область, а в точку, т.е. такой, что сама точка, в которую делается прогноз, не участвует в нем. Статистика L в этом случае рассчитывается на основе выражения (12).

$$L = \sum_{\bar{j}}\sum_{\bar{k}\in G_0} s_{\bar{j}} V^{-1}_{\text{т}\bar{j}\bar{k}}\left(z_{\bar{k}} - \hat{x}_{\text{т}\bar{k}}\right), \tag{12}$$

где $\hat{x}_{\text{т}\bar{k}}$ – оптимальный прогноз, сделанный на основе всех наблюдений, кроме $z_{\bar{k}}$; $V_{\text{т}\bar{j}\bar{k}} = P_{\text{т}\bar{j}\bar{k}} + \sigma_0^2 E_{\bar{j}\bar{k}}$ – ковариационная матрица наблюдений; $P_{\text{т}\bar{j}\bar{k}}$ – ковариационная матрица ошибок прогнозирования.

Процедура (12) требует фильтрации изображения, расчета ковариационной матрицы ошибок оценивания и взвешенного суммирования. Вероятность принятия ошибочного решения при заданном пороге λ_0 и выбранном допустимом уровне вероятности ложной тревоги можно найти по формуле (13) [18]:

$$P_F = \int_{\lambda_0}^{\infty} \omega(L \mid H_0)\,\mathrm{d}L = 0{,}5 - \Phi_0\left(\frac{\lambda_0 - M\{L \mid H_0\}}{\sqrt{D\{L \mid H_0\}}}\right), \quad (13)$$

где $M\{L \mid H_0\}$ и $D\{L \mid H_0\}$ – математическое ожидание и дисперсия отношения правдоподобия при условии отсутствия сигнала; $\Phi_0(\cdot)$ – функция Лапласа.

Соответственно, вероятность правильного обнаружения определяется формулой (14):

$$P_D = \int_{\lambda_0}^{\infty} \omega(L \mid H_1)\,\mathrm{d}L = 0{,}5 + \Phi_0\left(\frac{M\{L \mid H_1\} - \lambda_0}{\sqrt{D\{L \mid H_1\}}}\right). \quad (14)$$

При использовании приведенных соотношений нужно помнить, что они справедливы в предположении гауссовости ПРВ яркостей обрабатываемых изображений как в отсутствие сигнала, так и при его наличии.

Как правило, реальные изображения, например, спутниковые снимки, отличает существенная неоднородность фона. В этом случае параметры фона, в частности, ковариационную матрицу необходимо оценивать для каждого участка изображения. Однако для реальных изображений это требует больших вычислительных затрат. Использование же «усредненной» ковариационной матрицы, полученной в результате оценивания всего изображения, приводит к низкой достоверности обнаружения. Для решения данной задачи в работах [59 – 61] предлагается производить предварительную сегментацию изображения. Затем неоднородность фона выравнивается за счет приведения средних яркостей в сегментированных областях к средней яркости всего изображения, после чего проводят обнаружение объектов на таком псевдооднородном фоне. Явным достоинством описанного подхода является тот факт, что корректно выполненная процедура сегментации уже сама по себе позволяет выявить области изображения, потенциально являющиеся объектами интереса. Однако процедура сегментации сама по себе представляет сложную задачу [37, 60, 61]. Используемые для этого методы различны по своей эффективности и сложности [37, 51, 59, 62 – 65]. Некоторые основаны на вычислении вероятностей, с которыми полученные сегменты относятся к тому или иному классу [66], другие – на нахождении расстояния до среднего значения класса и порога принятия решения на основе среднеквадратичного отклонения [67]. Это может быть расстояние Махаланобиса или другое вероятностное расстояние [68]. В некоторых подходах характеристики для принятия решения базируются на схемах ранжирования данных [69].

Другим подходом к преодолению проблем, связанных с детектированием объектов на фоне пространственно неоднородных изображений, является предварительное использование процедур нелинейной фильтрации. В таком случае процесс обнаружения распадается на два этапа. На первом выполняется фильтрация изображения, а на втором – расчет некоторой статистики и сравнение ее с пороговым значением. В частности, такая двухэтапная процедура является очевидным результатом применения обнаружителя (12). При этом чем меньше будет ошибка фильтрации изображения, тем лучше будут окончательные результаты по обнаружению объектов. Для примера на рис. 7 приведено спутниковое изображение, содержащее аномалии круглой формы диаметром 10 пикселей на участке с лесным массивом (круг №1) и на участке сельскохозяйственного назначения (круг №2). На рис. 8 представлены рассчитанные кривые зависимости вероятности правильного обнаружения P_D от отношения сигнал/шум q при фиксированной вероятности ложной тревоги $P_F = 10^{-3}$. Различия в эффективности обнаружения одной и той же аномалии в данном случае определяются погрешностью фильтрации разных участков изображения. Линия 1 показывает характеристики обнаружения на лесном участке, линия 2 – на сельскохозяйственном.

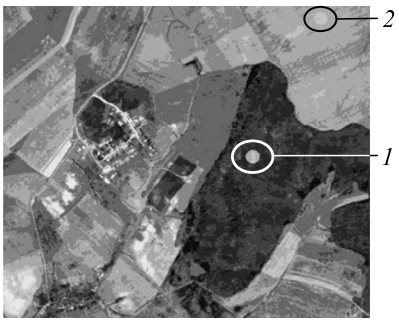

Рис. 7. Пример аномалий круглой формы на спутниковом изображении

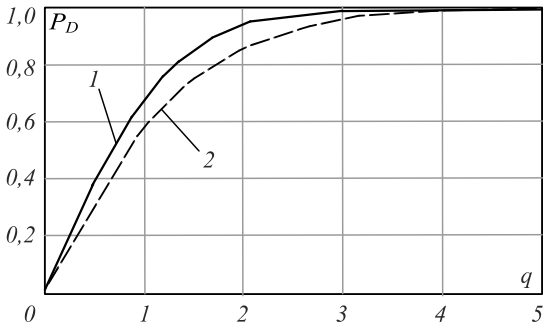

Рис. 8. Зависимость вероятности правильного обнаружения от отношения сигнал/шум

Опыт показывает [70, 71], что при обработке пространственно неоднородных изображений хорошие результаты демонстрируют алгоритмы, построенные на основе дважды стохастической фильтрации [72]. В этом случае изображение рассматривается как реализация некоторой дважды стохастической модели [73],

параметры которой сами по себе являются реализациями некоторых вспомогательных случайных полей. Такое представление позволяет, с одной стороны, успешно имитировать пространственно неоднородные изображения, а с другой (при наложении некоторых ограничений [71]) – строить эффективные процедуры обработки таких изображений, в том числе и обнаружения.

На рис. 9 показана возможность описания изображений с помощью дважды стохастической модели. В частности, рис. 9*а* иллюстрирует исходное изображение, а рис. 9*б* – представление его с помощью математической модели дважды стохастического случайного поля.

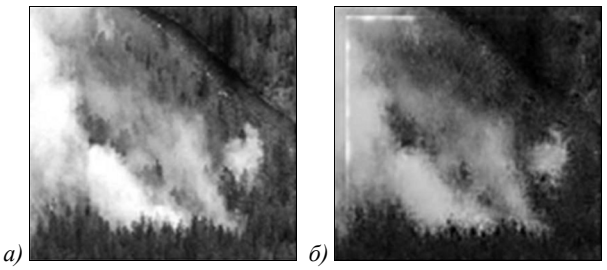

Рис. 9. Формирование изображения на основе дважды стохастической модели

На рис. 10 приведены зависимости дисперсии погрешностей фильтрации изображения D_e от дисперсии шума D_n при единичной дисперсии сигнала. Здесь кривая 1 соответствует результатам применения векторного фильтра; кривая 2 – дискретного фильтра Винера, кривая 3 – векторного фильтра Калмана с обратным ходом, кривая 4 – дважды стохастического фильтра [74].

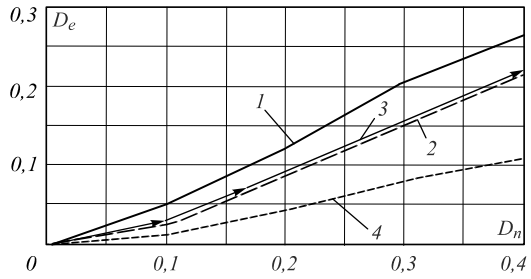

Рис. 10. Зависимость дисперсии погрешности фильтрации изображения от дисперсии шума

Анализ представленных на рис. 10 кривых показывает, что по дисперсии фильтрации дважды стохастический фильтр выигрывает более чем в два раза даже у лучших алгоритмов, не учитывающих изменяющиеся свойства изображений. Более того, даже в ситуации, когда остальные фильтры применяются уже к сегментированному изображению, выигрыш составляет порядка 15 – 20 %. Это объясняется возможностью адаптивной подстройки параметров дважды стохастического фильтра под изменяющиеся вероятностные свойства изображения.

Улучшение качества фильтрации изображений существенно повышает и вероятность обнаружения объектов. Подтверждением повышения эффективности обнаружения служит рис. 11, где на рис. 11*а* показан пример зашумленного изображения (отношение сигнал шум $q = 0{,}5$), содержащего малоразмерные объекты, а на рис. 11*б* – результаты дважды стохастической фильтрации и оценка вероятности правильного обнаружения, полученные после применения дважды стохастического фильтра (слева) и для сравнения после применения линейных фильтров (сверху) для одних и тех же областей сигналов.

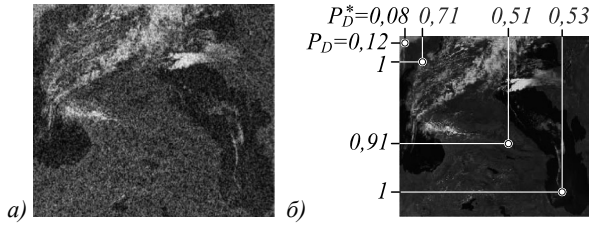

Рис. 11. Зашумлённое (а) и отфильтрованное (б) изображения с вероятностями правильного обнаружения

Анализ полученных результатов показывает, что алгоритм фильтрации на базе дважды стохастических моделей в абсолютном большинстве случаев обработки реального материала обеспечивает существенно лучшие характеристики, чем обнаружитель на базе линейных моделей.

Еще один подход к обработке неоднородного фона – применение адаптивных процедур [75]. Адаптация заключается в том, что перед обнаружением компенсация фона осуществляется с учетом его коррелированности. Так, в работе [76] применена безыдентификационная псевдоградиентная компенсация фона. При этом из очередного наблюдения $z_{\bar{j}}$ вычитается некий прогноз $\hat{z}_{\bar{j}}$, сделанный на основе соседних наблюдений (не содержащих $z_{\bar{j}}$). Прогноз формируется вектором параметров $\bar{\alpha}$, которые адаптируются к характеристикам изображения. Подбирается такой вектор параметров $\hat{\bar{\alpha}}$, который обеспечивает экстремум некоторой целевой функции $J(\hat{\bar{\alpha}})$, например, минимум дисперсии ошибки прогноза, аналитическое решение для которого найти не удается. При этом вектор параметров на очередном шаге формируется на основе параметров предыдущего шага и вектора обучения $\bar{\lambda}_{\bar{j}}$. Процесс настройки параметров описывается итерационной процедурой (15).

$$\hat{\bar{\alpha}}_i = \hat{\bar{\alpha}}_{i-1} - \\ -\bar{\lambda}_{\bar{j}}\,\mathrm{sign}\left(z_{\bar{j}i} - \hat{z}_{\bar{j}i-1}(\hat{\bar{\alpha}}_{i-1})\right)\left[\frac{\partial \hat{z}_{\bar{j}i-1}(z_{\bar{j}}, \hat{\bar{\alpha}}_{i-1})}{\partial \hat{\bar{\alpha}}_{i-1}}\right]. \qquad (15)$$

При оптимальных значениях оценок $\hat{\bar{\alpha}}$ параметров и $i \to \infty$ математическое ожидание ошибок прогноза также будет стремиться к 0. На практике число итераций ограничено и может составлять от единиц до сотен.

В работах [77 – 79] для поиска аномалий на многомерных изображениях применен аппарат теории нечетких множеств и генетические алгоритмы.

Нечеткое множество $\tilde{A}$, заданное на универсальном множестве X, определяется как множество пар $\tilde{A}=\{x_i, \mu_{\tilde{A}}(x_i)\}$, $i=1, 2, ..., n$, где $x_i \in X$ – элемент универсального множества, а $\mu_{\tilde{A}}(x_i) \in [0;1]$ – значение функции, определяющей степень принадлежности элемента x_i к нечеткому множеству $\tilde{A}$. После преобразования изображения в нечеткую область к полученному фаззифицированному (приведенному ко множеству) изображению можно применять известные нечеткие алгоритмы, в том числе и для решения задач обнаружения [77].

Идея генетических алгоритмов [78, 79] состоит в автоматизированном построении квазиоптимальных процедур обнаружения и идентификации объекта на изображении по его эталонному изображению с использованием библиотеки изображений обучающей выборки. При этом параметры обнаружителя итерационно подстраиваются, обеспечивая в рамках обучающей выборки квазиоптимальные характеристики обработки. Подобные процедуры обнаружения и идентификации в чем-то идеологически близки к рассматриваемым далее нейросетевым процедурам. Они легко обобщаются на случай увеличения размерности изображений, как и модели случайных полей. Тем не менее генетические алгоритмы имеют и ряд ограничений применения. В частности, корректный анализ эффективности генетических алгоритмов в настоящее время наталкивается на труднопреодолимые вычислительные сложности.

Для задач обнаружения и идентификации объектов на изображениях представляют также интерес алгоритмы, основанные на модельно-ориентированной методологии разработки алгоритмического и программного обеспечения [80]. При таком подходе по входному структурному описанию объекта поиска осуществляется автоматическое построение процедур его обнаружения, которые реализуют алгоритмы из заданного набора. В качестве базовых наборов алгоритмов для обнаружения объектов используются алгоритмы, основанные на иерархических нерекурсивных структурно-вероятностных и рекурсивных структурно-вероятностных моделях.

Для ситуаций, когда известен приблизительный разброс яркостей обнаруживаемого объекта (в одном или нескольких каналах мультиспектрального изображения), может быть эффективен и более простой подход – обнаружение по яркостному порогу [14, 19, 68, 81]. Разброс яркостей можно оценить с помощью построения гистограммы или определения хода спектральных кривых изображения объекта, полученных по его тестовым точкам. Затем задаются пороги, в пределах которых может колебаться яркость объекта интереса.

Существует также подход к улучшению характеристик систем обнаружения за счет внедрения в изображение дополнительных, искусственно сгенерированных аномалий. В частности, в работе [82] показано, как искусственные аномалии могут помочь в решении задачи обнаружения границ объекта. Но такие методы могут быть эффективными, если существуют надежные механизмы выделения искусственных аномалий на фоне объектов интереса.

В некоторых задачах высокую эффективность обнаружения может обеспечить учет цветовой палитры объекта, которую также можно оценить по канальным гистограммам эталонных изображений объекта [83]. Но понятно, что область применения такого подхода весьма узкая. Быстрое и эффективное в плане метрик обнаружения применение данного алгоритма возможно для объектов, в значительной степени характеризующихся их цветом.

Большое количество алгоритмов обнаружения в последние годы предложено в области анализа спутниковых многозональных и гиперспектральных изображений. В частности, распространение при обнаружении небольшого числа аномальных объектов на достаточно большом изображении получил алгоритм RXD [84], названный в честь своих авторов I.S. Reed и X.Yu. Он основан на анализе средних спектральных сигнатур всего изображения и сигнатур, рассчитанных для отдельных участков изображения. Однако для изображений с более сложной структурой, у которых сигнатуры фона на разных участках могут значительно отличаться, эффективность такого подхода снижается. Для преодоления этого недостатка предложены различные модификации RXD-алгоритма [85, 86], основанные на предварительной обработке многозональных изображений. Например, с использованием спектрального разложения и метода главных компонент упрощают описание изображения, что повышает эффективность обнаружения. Еще одна модификация заключается в проведении пространственной сегментации изображений, на основе которой осуществляется переход к псевдооднородным изображениям. Также к способам предобработки относится использование двух скользящих окон [85]: большого – для выделения однородных участков, а маленького – для поиска аномалий внутри таких участков.

Отметим кратко еще несколько подходов. В работах [87, 88] изображение разбивается на атрибуты, которые и являются элементами модели при обнаружении. К атрибутам, в частности, можно отнести точки (углы, соединения линий, точки большого градиента яркости, центр тяжести области, концы линий, точки экстремальных значений признаков), линии (прямые или криволинейные структуры, границы областей), области (сегментированные области, специфические формы) и т.д. Базовые критерии, которые могут быть использованы при выборе характерных черт изображений для задачи обнаружения, рассмотрены в [88].

Прежде чем переходить к методам на базе современных глубоких искусственных нейронных сетей (ИНС), также следует отметить подходы «зрения на

основе модели» [89]. Указанные подходы уже 15 – 20 лет назад позволяли получать показатели качества, сопоставимые с показателями современных ИНС, но уступали по «технологичности» подхода. Например, можно отметить обнаружитель, предложенный Виолой и Джонсом в 2001 г. [90], позволявший вести обработку изображений в реальном времени. Несмотря на то, что алгоритм был способен распознавать различные объекты, основное применение он нашел в системах идентификации лиц, поскольку базировался на простых каскадах Хаара [91], позволяющих эффективно и быстро решать данную задачу.

Необходимо также упомянуть и другие подходы, которые предшествовали широкому распространению ИНС. Многие из них имеют непосредственное отношение к задаче идентификации лиц, однако являются в то же время достаточно хорошими дескрипторами для задач обнаружения и распознавания. Общая классификация таких алгоритмов подразумевает выделение трех основных подходов:

1) Структурный подход: все объекты описываются как системы, включающие в себя большое количество взаимосвязанных элементов.

К основным методам структурного подхода можно отнести масштабно-инвариантную трансформацию признаков (scale-invariant feature transform, SIFT) [92], в результате которой сначала обнаруживаются такие признаки, затем происходит их сопоставление и индексация и, наконец, решается задача идентификации.

Также сюда относится метод устойчивых ускоренных признаков (Speeded Up Robust Features, SURF) [93]. Данный детектор / дескриптор инвариантен не только к масштабу, но и к вращению особых точек, которые свойственны, например, лицу конкретного человека.

Еще одним представителем является метод бинарных устойчивых независимых элементарных функций (binary robust independent elementary features, BRIEF) [94]. BRIEF представляет собой дескриптор универсальной точки, который можно комбинировать с произвольными детекторами. Он устойчив к типичным классам фотометрических и геометрических преобразований изображений.

Метод локальных бинарных шаблонов (local binary pattern, LBP) [95]. Данный метод широко используется для распознавания лиц, выражений лица, сегментации текстур, а также их классификации. LBP описывает окрестности пикселей в двоичном представлении.

Также широкое распространение получил метод гистограммы ориентированных градиентов (histogram of oriented gradients, HOG) [96]. Гистограмма направленных градиентов – один из лучших дескрипторов, используемых при обнаружении объектов. Метод HOG помогает описать форму объекта на изображении используя распределение градиентов интенсивности или направлением краев. Процесс реализации данной техники заключается в разделении целого изображения на области; построении гистограммы направления краев для пикселей или направлений градиентов для каждой из областей; после чего комбинация полученных гистограмм используется для извлечения признаков объекта. Для расчёта значений градиентов чаще всего применяется одномерная дифференцирующая маска в горизонтальном и вертикальном направлениях с фильтрующим ядром [–1, 0, 1] в соответствии с выражением (16).

$$\begin{aligned} G_x(x,y) &= I(x+1,y) - I(x-1,y), \\ G_y(x,y) &= I(x,y+1) - I(x,y-1), \end{aligned} \tag{16}$$

где $I(x,y)$ – значение пикселя в точке (x,y);
$G_x(x,y)$ и $G_y(x,y)$ – амплитуды горизонтального и вертикального градиента соответственно.

Величина градиента и ориентация каждого пикселя (x,y) вычисляются в соответствии с выражением (17).

$$\begin{aligned} G(x,y) &= \sqrt{G_x^2(x,y) + G_y^2(x,y)}, \\ \theta(x,y) &= a\tan\left(\frac{G_y(x,y)}{G_x(x,y)}\right). \end{aligned} \tag{17}$$

2) Холистический подход описывает объекты как единое целое, проецируя все изображение в меньшее подпространство или в плоскость корреляции, включает в себя такие линейные методы, как метод главных компонент (principal component analysis, PCA), линейный дискриминантный анализ (linear discriminant analysis, LDA), Eigenfaces, и нелинейный метод главных компонент KPCA (kernel PCA), сверточные нейронные сети (convolutional neural network, CNN), машина опорных векторов (support vector machine, SVM) [97].

3) Гибридный (кластерный) подход совмещает структурный и холистический подходы для повышения точности распознавания.

Кроме того, представления на основе деталей широко используются для задач визуального распознавания. В частности, модели деформируемых деталей (Deformable Parts Models, DPM) [98] были особенно полезны для определения общих категорий объектов. DPM обновляют модели графической структуры [99], которые восходят к 1970-м годам, с помощью современных функций изображения и алгоритмов машинного обучения.

При использовании математических моделей для обнаружения аномалий или разного рода объектов особое внимание необходимо уделять анализу их применимости для конкретной задачи. Можно получить отличные результаты на искусственных данных, но при обработке реальных изображений многое зависит от адекватности используемых моделей этим изображениям. Во многом этим объясняется отсут-

ствие универсальных алгоритмов обнаружения на базе математических моделей изображений.

3. Глубокое обучение и сверточные нейронные сети в задачах обнаружения объектов на изображениях

Появление методов глубокого обучения и сверточных нейронных сетей позволило для обучения алгоритмов обнаружения применять вместо моделей изображений реальные снимки. Уже в прошлом остались те времена, когда одной из принципиальных задач компьютерного зрения было различение изображений кошек и собак [100]. Сегодня нейронные сети решают куда более сложные задачи. К примеру, сеть Mask R-CNN [101] позволяет выделять контуры множества представителей объектов одного или разных типов. Таким образом, решается задача объектовой сегментации, сочетающая в себе задачи обнаружения, классификации и сегментации.

Рассмотрим подробнее CNN, состоящие из слоев свертки, субдискретизации (пулинга) и слоев с полной связью (полносвязных). Назначение сверточного слоя – объединение яркости пикселей в локальной окрестности для последующего выделения общих признаков, характерных для изображения. Процедура свертки выполняется скользящим квадратным окном небольшого размера (3×3, 5×5, 7×7 … пикселей), так называемым ядром свертки (kernel). Оно реализует взвешенное суммирование пикселей внутри локальной области (которое и «объединяет» яркости), а полученная сумма характеризует признак, соответствующий локальной области изображения, попавшей в ядро свертки. В результате движения такого окна по всему изображению (например, слева направо и сверху вниз) происходит получение матрицы или карты признаков, в которой в отдельных точках будут представлены взвешенные суммы пикселей из локальных окрестностей. Если в архитектуре сети отсутствует слой субдискретизации, известный также как слой пулинга или объединения, то на следующий сверточный слой попадает уже карта признаков, полученная на предыдущем слое. В противном случае на такой слой подается карта признаков. Процесс свертки иллюстрируется рис. 12.

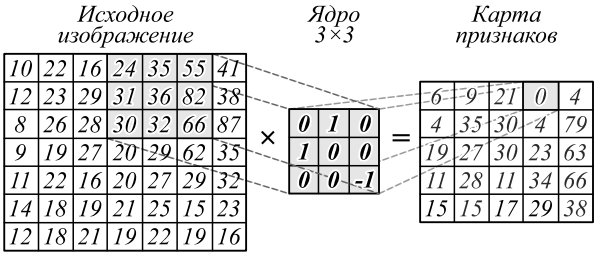

Рис. 12. Пример формирования карты признаков

Для более полного извлечения признаков входное изображение и последующие карты признаков могут формироваться с использованием разных ядер. Тогда на выход сверточного слоя поступает несколько карт признаков, что позволяет более детально извлекать информацию из объектов на изображении. В случае обучения с учителем, т.е. когда известны «правильные» выходные результаты работы сети, весовые коэффициенты в ядрах свертки формируются во время обучения с использованием метода обратного распространения ошибки [102, 103], когда начиная с последнего слоя сети (уже не сверточного, а полносвязного) веса корректируются на основании соответствия текущего и «правильного» результатов на выходе.

Ядро свертки формирует новое значение в своем геометрическом центре, поэтому размер карты признаков на выходе будет меньше размера на входе. Однако существуют и алгоритмы, обеспечивающие одинаковый размер на входе и выходе. В них для «пропадающих» краев карты признаков формируются так называемые значения по внутренним отступам (paddings-значения) [104], позволяющие сохранить тот же размер после свертки. Может задаваться также и параметр шага, с которым перемещается окно. Чем больше шаг сверточного ядра по базовым осям изображения, тем меньше по размерам выходная карта признаков и точность их выделения.

Более того, формирование выходного значения ядра может реализовываться не только взвешенным суммированием, но и по другим правилам, например выбором максимальной яркости. Полученный таким образом слой, получивший название «MaxPooling» [105], используется в дополнение к сверточным слоям. Ядро максимизации применяется, как правило, к непересекающимся областям изображения, как показано на рис. 13, и позволяет заменить группу признаков одним значением. Понятно, что оно тоже должно быть небольшого размера, иначе извлеченные признаки могут быть не достаточно информативны.

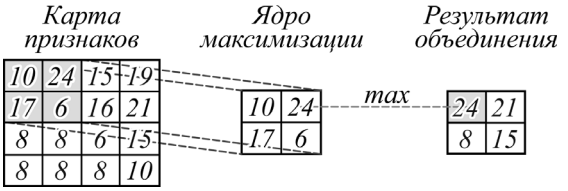

Рис. 13. Пример работы слоя MaxPooling

Рассмотрим кратко развитие сверточных нейронных сетей для решения задач обнаружения.

Сеть R-CNN (сокращение от «Region-based Convolutional Network» – «сверточная сеть на базе регионов») [25] была разработана в университете Калифорнии UC Berkeley для задачи обнаружения объектов [25]. На тот момент сверточные сети уже хорошо зарекомендовали себя в задачах распознавания. Сеть R-CNN позволила решать данную задачу не на всем изображении, а на предварительно выделенных областях (регионах), где могут присутствовать объекты. Для выделения областей был выбран метод селективного поиска [106], а для распознавания – архитектура

CaffeNet (являющаяся вариантом сети AlexNet) [24], направленная на распознавание объектов 1000 классов и обученная на наборе изображений ImageNet [107]. Несмотря на то, что базовые распознаватели обучены под такое большое число классов, детектор R-CNN применялся для обнаружения объектов гораздо меньшего числа классов. В частности, известны версии для 20 и 200 классов объектов. Авторами был заменен и дообучен последний слой сети AlexNet, для чего был добавлен класс фонового изображения. После селективного выбора формировалось порядка 2000 прямоугольных областей разной площади и протяженности по базовым осям изображения. Выбранные регионы перед распознаванием предварительно модифицировались под размеры 227×227 пикселей, поскольку с изображениями такого формата работает сеть CaffeNet, которая формирует для них вектор признаков, содержащий 4096 элементов. Классификация осуществляется с использованием метода опорных векторов [108], предполагающего формирование N линейных векторов (по числу классов объектов). Опорные вектора используются для установления факта нахождения объектов, принадлежащих одному из N классов (без учета фона). Такая структура позволила минимизировать матричные операции. Для всех 2000 регионов, содержащих по 4096 признаков, формируется матрица размером 2000×4096, которая умножается на матрицу, содержащую эталонные признаки для каждого класса объектов размером $4096 \times N$ (матрицу с весами опорных векторов). На выходе получается матрица размером $2000 \times N$, которая после бинаризации показывает, какие из классов объектов присутствуют в каждом регионе. Эта процедура может быть проиллюстрирована схемой (рис. 14), схожей с приведенной в работе [25].

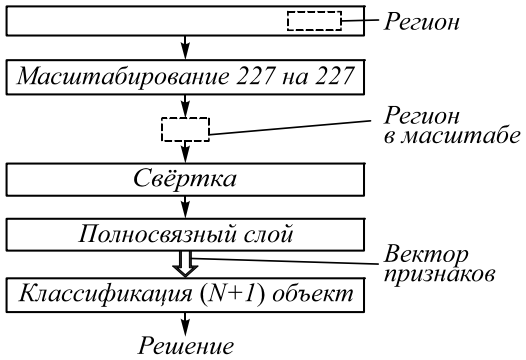

Рис. 14. Схема сети R-CNN

Регионы могут содержать не только целые изображения объектов, но и лишь их часть. Дополнительно для принятия решения о том, содержит регион объект или нет, используется метрика *IoU*, более подробно рассмотренная ранее. Ее применение позволяет устранить избыточность регионов, содержащих объекты одного класса. Так, например, если *IoU* между исследуемым регионом и регионом, содержащим максимальную вероятность для данного класса, ниже выбранного порога, то данный регион может быть исключен. Для повышения точности определения *IoU* был разработан метод, позволяющий скорректировать параметры ограничивающих прямоугольников – Bounding-box regression [25]. Его идея состоит в том, что после распознавания изображения региона с использованием линейной регрессии на основе признаков корректируются координаты центра, ширина и высота прямоугольника.

Таким образом, выделим основные этапы работы сети R-CNN:

- селективный выбор регионов для анализа наличия в них определенных объектов;
- деформация региона под заданный размер сети CaffeNet;
- извлечение вектора признаков региона;
- бинарная классификация вектора признаков региона для каждого из возможных объектов с помощью метода опорных векторов;
- уточнение местоположения ограничивающего прямоугольника региона с использованием линейной регрессии.

Для объектов из базы ImageNet архитектура сети R-CNN обеспечивала достаточно высокую точность и полноту, но производительность ее была относительно небольшой, что ограничивало использование более глубоких сетей, например, сети с архитектурой VGG16 (сокращение от «Visual Geometry Group») [109], имеющей 16 сверточных слоев. Кроме того, методы опорных векторов и регрессионного прогнозирования ограничивающего прямоугольника требовали больших объемов памяти. Поэтому авторами алгоритма R-CNN в 2015 году была предложена его более быстрая версия Fast R-CNN [110]. Для нее использовалось два подхода:

- выполнять извлечение признаков не для каждого региона на изображении, а для всего изображения с последующим наложением границ регионов на карту признаков;
- проводить одновременное обучение всех трех процедур, используемых в алгоритме: свертки, формирования опорных векторов и линейной регрессии.

Для преобразования признаков из разных регионов к заданным размерам был предложен слой объединения регионов интереса (Region of Interest Pooling). Для этого область региона делится на сетку с ячейками размером $H/h \times W/w$, где H и W – размеры региона, а h и w – размеры ядра объединения. В каждой ячейке выбирается максимальный элемент. Отдельного применения метода опорных векторов в новом алгоритме больше не было, а отобранные признаки передавались на полносвязный слой и затем для параллельной обработки – на слой классификации с $N+1$ выходом (с учетом фона) и слой регрессии ограничивающего прямоугольника (Bounding-Box

Regression). Общая архитектура сети поясняется схемой на рис. 15.

Однако в сети Fast R-CNN также оставалось узкое место: низкая производительность алгоритма селективного выбора регионов интереса. В 2015 г. с появлением архитектуры Faster R-CNN [111] этот этап удалось значительно ускорить. В данном обнаружителе регионы вычисляются на основе полученной карты признаков. Это потребовало добавления еще одного модуля, который анализирует карту признаков и на ее основе предлагает регионы интереса. Данный модуль получил название RPN (сокращение от «Region Proposal Network» – сеть предложений регионов), т.е. специальная сеть, предлагающая регионы для исследования. Архитектура Faster R-CNN приведена на рис. 16.

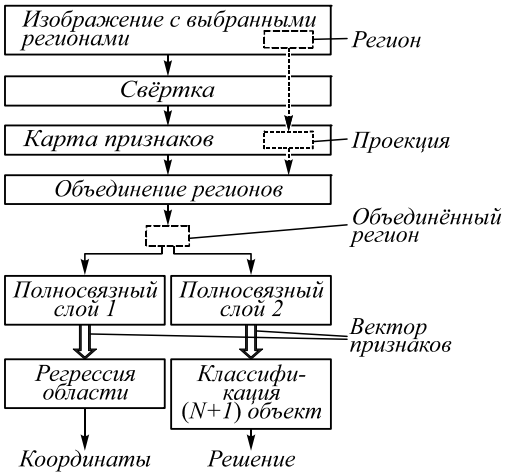

Рис. 15. Схема сети Fast R-CNN

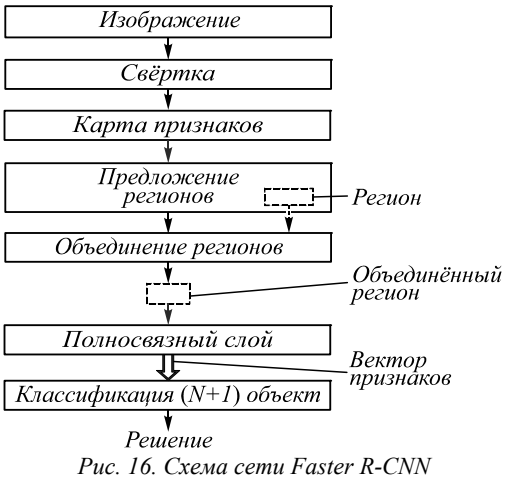

Рис. 16. Схема сети Faster R-CNN

Модуль RPN обрабатывает извлеченные признаки всего изображения с применением простой нейронной сети. При обработке используется окно небольшого размера, например, 3×3 признака. Выходы с такой нейросети передаются параллельно для обработки двумя полносвязными слоями. Первый слой предназначен для оценки линейной регрессии геометрических параметров региона, а второй – отвечает за непосредственную классификацию объектов в регионе. Выходами таких слоев являются некие якоря (ограничивающие прямоугольники) с разными размерами для каждого положения скользящего окна. Первый слой для каждого якоря определяет 4 координаты для уточнения положения области объекта. Второй слой выдает два числа, определяющие вероятность присутствия или отсутствия объекта в регионе (рис. 17). Это позволяет исключить «неинтересные» регионы. Обучение обоих слоев происходит одновременно. Функция потерь задается в виде суммы функций потерь для каждого слоя с весовыми коэффициентами. Оставленные регионы передаются в блок обработки, реализованный в соответствии с архитектурой Fast R-CNN [110].

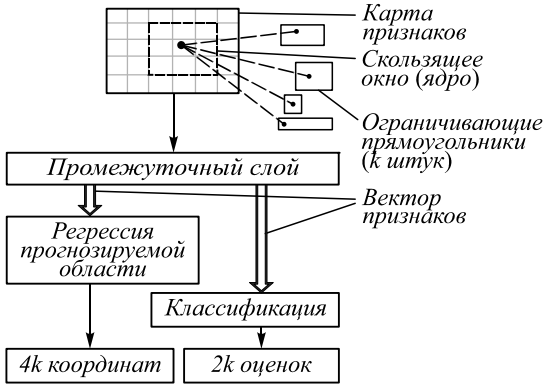

Рис. 17. Метод предложения регионов на базе якорей

Обучение сети Faster R-CNN происходит в несколько этапов:

1) Обучение модуля RPN для выявления регионов интереса.
2) По отобранным регионам происходит переобучение сети Fast R-CNN.
3) Обученная на этапе 2 сеть используется для задания весов в сети RPN. Следует отметить, что в этом случае замораживаются общие слои свертки, а перенастраиваются только слои, связанные с RPN.
4) С учетом зафиксированных слоев окончательно происходит настройка параметров сети Fast R-CNN.

Расширенную версию Faster R-CNN представляет архитектура Mask R-CNN [101]. Ее основное отличие заключается в том, что за счет добавления еще одного полносвязного слоя предсказывается маска, покрывающая объект, а не ограничивающий прямоугольник (регион). За счет этого сеть Mask R-CNN, помимо задач обнаружения и распознавания, решает задачу сегментации, поскольку позволяет выделять отдельные объекты одного класса на изображении как разные сегменты. Маска имеет простой вид – это бинарная матрица, в которой «1» показывает принадлежность пикселя к объекту, а «0» – не принадлежность.

Функция потерь в такой сети представляет собой сумму функций потерь для классификации, обнаружения и сегментации. Выделение масок происходит

без предварительного обнаружения в регионе. Выбирается маска, имеющая наибольшую вероятность в независимом классификаторе. Для каждого класса получаются свои бинарные маски. Пример работы сети Mask R-CNN представлен на рис. 18 [101].

Рис. 18. Пример результата работы сети Mask R-CNN

Важно отметить, что для архитектуры Fast R-CNN, где необходимо выполнить проекцию региона с исходного изображения на карту признаков, могут возникать сложности, связанные с получением такой проекции. Действительно, карта признаков, формируемая простой сверточной нейронной сетью, имеет фиксированный размер, который меньше размера исходного изображения, поэтому регион, содержащий на изображении целое число пикселей, не получается отобразить в пропорциональный регион карты признаков, который бы идеально накладывался на ячейки данной карты (регион с целочисленным количеством признаков). Для примера на рис. 19*а* приведены фрагмент региона (с серой полупрозрачной заливкой) и фрагмент сетки карты признаков (без заливки, элементы отделены толстыми линиями). Пунктирные линии ограничивают области на регионе, для которых необходимо выполнить объединение.

В сети Fast R-CNN проблема решалась округлением дробных значений координат до целых. Тогда для приведенной на рис. 19 ситуации получаем фрагмент сетки карты признаков, показанный на рис. 19*б*. Такой подход нормально работает при выделении ограничивающей рамки, но полученная таким путем маска получается неточной. Для повышения точности в сети Mask R-CNN не используется округление координат, все числа остаются действительными, а для вычисления значений признаков используется билинейная интерполяция по четырём ближайшим целочисленным точкам [101]. Для этого в каждой ячейке карты признаков строится дополнительная сетка. Для примера на рис. 19*в* она взята размером 2×2 и показана пунктирными линиями. Для каждой ячейки, не округляя координаты, с использованием билинейной интерполяции можно найти значение по 4 ближайшим значениям региона изображения. Затем по каждой ячейке карты признаков формируется слой Max-Pooling. Для рассматриваемого примера (рис. 19*в*) результат показан на рис. 19*г*. Такая процедура интерполяции маски получила название RoI Align (выравнивание регионов интереса).

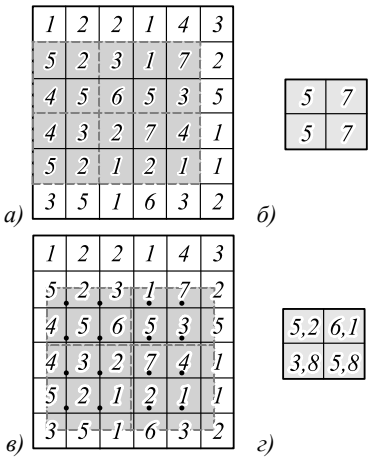

Рис. 19. Интерполяция региона

Помимо высоких результатов при детектировании объектов, сеть Mask R-CNN показала хорошие результаты при определении поз людей на изображении. Для этого сеть обучают так, чтобы она выдавала в масках только один «не нулевой» пиксель, местоположение которого соответствует опорным точкам изображения человека, таким как левые (правые) плечо, локоть, колено и т.п. По таким опорным точкам легко можно оценить каркас позиции человека. Сеть обучается выдавать по одной «однопиксельной» маске для каждого типа опорной точки. Пример такой работы сети, взятый из [112], приведен на рис. 20.

Однако задачи семантической сегментации и выделения скелета человека на изображении лишь частично пересекаются с задачами обнаружения объектов и требуют для оценки качества своих метрик, анализ которых выходит за пределы данной статьи.

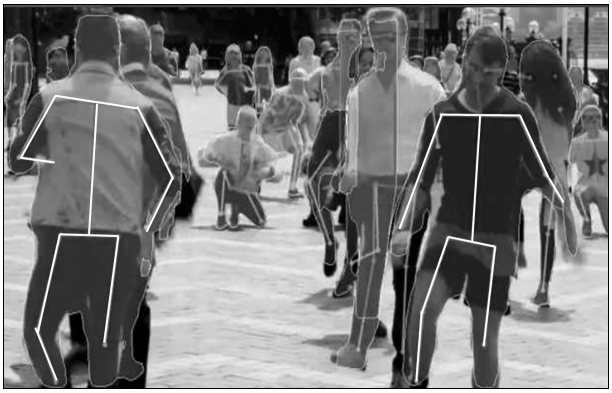

Рис. 20. Пример построения каркаса позиции человека

Еще одним классом нейронных сетей, используемым при обнаружении объектов, являются сети, которые одновременно за один проход формируют ограничивающие прямоугольники и предсказывают

класс объекта [113]. Примерами таких архитектур являются сети YOLO, SSD, RetinaNet и др.

Архитектура YOLO в настоящее время является одним из самых популярных детекторов не только для анализа изображений, но и для обработки видео в реальном времени. С появлением третьей версии этой архитектуры [114] значительно возросли точность и производительность алгоритмов обнаружения. В основе сети YOLOv3 лежит сеть Darknet-53 [114], но в отличие от селективных методов выбора региона, применяемых в семействе архитектур R-CNN, в сети YOLO входное изображение разделяется на квадратные области, для которых выполняется классификация. Для каждого квадрата изображения прогнозируются три ограничивающих прямоугольника и оценивается достоверность присутствия в них объектов. В YOLOv3 возможно предсказание объектов 80 различных классов.

Благодаря высокому быстродействию и точности обнаружения, архитектура YOLO в настоящее время продолжает бурно развиваться. Так, в апреле 2020 г. в свет вышла четвертая версия YOLOv4 [115], в которой для повышения точности модифицирована процедура аугментации (расширения) данных. Чуть позже появилась и архитектура YOLOv5 [116]. Релиз YOLOv5 [117] состоит из 5 различных моделей YOLO, которые отличаются размерами.

Основное преимущество YOLOv5 по сравнению с предыдущими версиями было получено вследствие того, что в качестве фреймворка разработки применен вычислительный пакет PyTorch, тогда как предшествующие версии использовали DarkNet. Сеть обеспечила практически трехкратное ускорение работы детектора, а интерполированная средняя точность для датасета BCCD (сокращение от «Blood Cell Countand Detection» – «подсчет и обнаружение клеток крови»), представляющего набор фотографий клеток крови, составила $mAP = 0{,}895$ [118]. При этом веса модели требуют объема памяти практически на порядок меньше, чем веса модели YOLOv4 такой же точности. В настоящее время данная сеть по метрике mAP уже превзошла сеть EfficientDet [119], архитектура которой также соответствует типу SSD, а сам обнаружитель EfficientDet построен на модели EfficientNet [120], распознающей объекты из датасета ImageNet, путем добавления слоя с взвешенной пирамидой признаков и слоя для предсказания положения ограничивающего прямоугольника.

Архитектуры SSD [29] и YOLO используют идею якорей – выделение квадратов из изображения и прогнозирование для них покрывающих прямоугольников, для каждого из которых уточняются координаты, определяется приоритетный класс объекта и уровень достоверности отнесения к этому классу. Но существуют также сети, в основе которых при обнаружении используется описание диагонали (или угла) ограничивающего прямоугольника, получившие название CornerNet [121]. Они позволяют описать ограничивающий прямоугольник с помощью двух параметров: левый верхний и нижний правый углы. Однако данная идея не имеет существенных отличий от идеи построения ограничивающего прямоугольника с соответствующими шириной и высотой.

Осложняющим обстоятельством при обучении сетей, работающих в один проход, является то, что большинство обучающих примеров содержит фон изображения, что приводит к ухудшению качества при обнаружении объектов. Для минимизации влияния фона в архитектуре RetinaNet введена функция коррекции потерь Focal Loss [122]. Основу сети RetinaNet составляет сеть FPN (сокращение от «Feature Pyramid Network» – «функциональная сеть пирамид») [123] и модель ResNet [107].

Также стоит отметить, что сети архитектуры YOLO выгодно отличаются именно в плане производительности, т.е. характеризуются достаточно большим числом кадров, обрабатываемых в единицу времени [124].

В табл. 1 приводятся результаты сравнения различных алгоритмов по средней точности обнаружения (Box AP, %) с учетом корректности формирования ограничивающих прямоугольников для датасета COCO [125] в равных условиях обучения.

Табл. 1. Сравнительные характеристики нейросетевых обнаружителей

Обнаружитель	**Box AP, %**	**Дополнительные данные при обучении**
YOLOv5 [116]	59,2	Нет
Swin-L [126]	58,7	Нет
CenterNet2 [127]	56,4	Да
YOLOv4-P7 [128]	56	Нет
EfficientDet-D7 [119]	53,7	Да
Cascade Mask R-CNN [129]	53,3	Да
RetinaNet [130]	52,1	Да
Mask R-CNN [131]	46,1	Нет
Cascade R-CNN-FPN [132]	45,9	Нет
Faster R-CNN [133]	43,9	Нет
Fast R-CNN [134]	40,1	Нет
SSD512 [29]	28,8	Нет
YOLO v2 + Darknet-19 [135]	21,6	Нет

Следует отметить, что алгоритмы сравниваются на одинаковых тестовых выборках из COCO, а обучение проходит при одинаковых гиперпараметрах, что делает сравнительные результаты адекватными.

Анализ представленных результатов показывает, что трансформерная архитектура Swin-L уже в настоящее время обеспечивает достаточно высокую эффективность решения задачи обнаружения объектов. Рассмотрим данную модель более подробно.

Предлагаемый Swin Transformer строит иерархические карты функций путем объединения участков изображения в более глубоких слоях и имеет линейную сложность вычислений для размера входного изображения из-за вычисления «самовнимания» [136] только в каждом локальном окне. Таким образом, он может слу-

жить основой общего назначения как для классификации изображений, так и для задач плотного распознавания. Напротив, предыдущие Vision Transformers [136] производили карты характеристик с одним низким разрешением и имели квадратичную вычислительную сложность для размера входного изображения из-за глобального вычисления собственного внимания.

Также при сравнении нейросетевых архитектур немаловажным является анализ самих моделей. В табл. 2 приводятся характеристики ряда нейросетевых обнаружителей.

Табл. 2. Сравнительный анализ архитектур

Обнаружитель	**Число слоев**	**Число параметров**
YOLOv5 [114]	191	7,5 M
Swin-L [124]	192	200 M
Faster R-CNN [131]	6+	20 K
Fast R-CNN [132]	6+	4 K
EfficientDet-D7 [117]	813	11 M
RetinaNet [128]	10	60 M

Из представленной таблицы видно, что архитектуры могут быть достаточно глубокими и сложными, однако самые простые архитектуры будут у сетей, для которых приоритетным является быстродействие.

Таким образом, можно отметить, что переломный момент в алгоритмах распознавания наступил с появлением нейронной сети AlexNet в 2012 г., что подтверждают результаты соревнований на наборе ImageNet, характеризующиеся процентом ошибочных распознаваний, с 2010 по 2014 гг., представленные на рис. 21.

Рис. 21. Ошибки распознавания по годам

Что касается обнаружителей, то результаты развития метрики средней точности *mAP* (на датасете COCO) по годам (с 2016 по 2021 гг.) представлены на рис. 22.

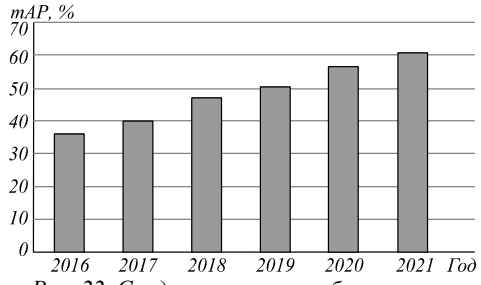

Рис. 22. Средняя точность обнаружения

Выше отмечалось, что на данный момент лучшей считается архитектура YOLOv5 [137]. Однако к моменту выхода статьи могут быть получены и новые результаты, например, на основе применения моделей трансформерного типа, широко распространенных при решении задач обработки естественного языка [138, 139]. В частности, уже получены первые результаты в области такого обнаружения [126, 136, 140]. Ещё одним перспективным направлением развития является применение для решения задач обнаружения мультимодальных данных [141]. Обработка данных такого рода, например, может находиться на стыке сферы компьютерного зрения и обработки естественного языка. Для примера на рис. 23 показана схема обработки изображения на основе вопроса, сформулированного на естественном языке, аналогичная примеру из [142].

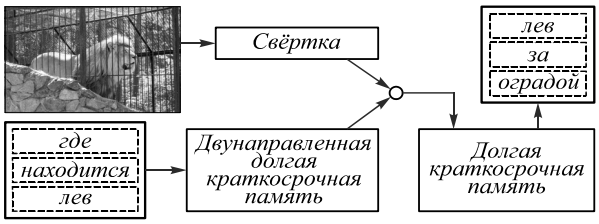

Рис. 23. Работа с мультимодальными данными

В представленном примере также присутствуют блоки BLSTM [143] (сокращение от «Bidirectional Long Short-Term Memory» – «Двунаправленная долгая краткосрочная память») и LSTM [144] (сокращение от «Long Short-Term Memory» – «долгая краткосрочная память»). Их работа связана с анализом и синтезом естественного языка, а обнаружение льва и ограды на изображении (рис. 23) осуществляется с помощью сверточной сети.

Таким образом, глубокое обучение не только предоставило ряд возможностей для разработки и модификации алгоритмов обнаружения объектов, но и остается на сегодняшний день очень бурно развивающейся сферой знаний, в том числе в задачах обработки изображений.

Наконец, кратко рассмотрим такое понятие, как бенчмарки (benchmarks) или методы тестирования алгоритмов, применительно к нейронным сетям. В работе [145] представлены результаты выполнения таких тестов для алгоритмов распознавания образов. При этом авторы уделяют внимание описанию аппаратного комплекса, который был использован для испытаний, а основными критериями, по которым сравниваются модели, являются:

1) Доля верных распознаваний (Accuracy).
2) Сложность модели (Model Complexity), описываемая числом обучаемых параметров.
3) Потребление памяти (Memory Usage), в том числе при различных размерах батчей.
4) Вычислительная сложность (Computational Complexity) как число умножений-сложений алгоритма.
5) Время инференса (Inference Time), характеризующееся временем, необходимым на обработку одного кадра изображения.

В качестве примера компаний, которые выполняют подобного рода бенчмарки, можно отметить DARPA или, например, ресурс https://eval.ai, https://kaggle.com, https://archive.ics.uci.edu/, на которых можно найти датасеты и сравнить различные алгоритмы.

Заключение

Задачи, связанные с обнаружением аномалий и объектов на изображениях и видеопоследовательностях, – это всего лишь относительно небольшая область задач компьютерного зрения. Но это очень важная и востребованная область. Разнообразие сфер ее приложений очень велико: обнаружение возгораний по спутниковым снимкам, дефектов железнодорожных путей, заболеваний по медицинским изображениям, состояния металла по металлографическим изображениям, лиц в системах контроля и наблюдения, видов заканчивающихся товаров в магазине и многое другое.

В методах обнаружения можно условно выделить два подхода: первый, основанный на математических моделях изображений и объектов интереса; и второй, основанный на обучении процедур обнаружения на библиотеках реальных снимков. К первому можно отнести использование моделей случайных полей, атрибутов изображения, нечетких множеств, моделей на основе генетических алгоритмов и другие методы. Ко второму – процедуры на основе нейронных сетей. В частности, в данной работе проанализировано развитие архитектур на базе сети R-CNN и сетей типа SSD, работающих за один проход. Нейронные сети способны обеспечивать высокие показатели в обнаружении, однако, в отличие от методов на основе моделей случайных полей, требуют значительных вычислительных ресурсов и огромного объема визуальной априорной информации об объектах. Тем не менее, эти проблемы сегодня разрешаются быстрым развитием аппаратного оборудования и методов аугментации данных. С другой стороны, обнаружение становится все более кастомизированным, т.е. заточенным под конкретные практические задачи.

Одновременно с развитием методов обнаружения развивались метрики и критерии качества обнаружения. Кроме «классических» вероятностей ошибок первого и второго рода при проверке гипотез о наличии объекта, появились точность и полнота обнаружения, пересечение по объединению, интерполированная средняя точность и другие метрики.

Также в статье были рассмотрены подходы, предшествующие непосредственно искусственным нейронным сетям, проанализирована сложность моделей и представлены типовые бенчмарки. Кроме того, отмечены ресурсы и организации, которые проводят такие бенчмарки.

Проблема разработки все более эффективных алгоритмов обнаружения остается актуальной уже не одно десятилетие, и впереди нас ждет еще много новых интересных решений, в частности в области применения в обнаружении трансформерных архитектур нейронных сетей и подходов к описанию мультимодального мира.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-17-50020 и частично проекта №19-29-09048.

References

[1] Bagautdinov RS, KopenkovVN, Myshkin VN, Sergeev VV, Tribunsky SA. Study of the applicability of satellite imagery to detecting archeological objects. Computer Optics 2015; 39(3): 439-444. DOI: 10.18287/0134-2452-2015-39-3-439-444.

[2] Andriyanov NA, Vasil'ev KK, Dement'ev VE. Investigation of filtering and objects detection algorithms for a multizone image sequence. ISPRS Archives 2019; XLII-2/W12: 7-10. DOI: 10.5194/isprs-archives-XLII-2-W12-7-2019.

[3] Attard L, Farrugia R. Vision based surveillance system. 2011 IEEE EUROCON – Int Conf on Computer as a Tool 2011; 1: 1-4. DOI: 10.1109/EUROCON.2011.5929144.

[4] Prati A, Shan C, Wang K. Sensors, vision and networks: From video surveillance to activity recognition and health monitoring. J Ambient Intell Smart Environ 2019; 11(1): 5-22. DOI: 10.3233/AIS-180510.

[5] Raghu M, Zhang C, Kleinberg J, Bengio S. Transfusion: Understanding transfer learning for medical imaging. Proc 33rd Conf on Neural Information Processing Systems (NeurIPS) 2019; 1: 1-22.

[6] Mikhaylichenko AA, Demyanenko YaM. Detection of the bone contours of the knee joints on medical X-ray images. Computer Optics 2019; 43(3): 455-463. DOI: 10.18287/2412-6179-2019-43-3-455-463.

[7] Zherdev DA, Minaev EY, Procudin VV, Fursov VA. Object recognition using real and modelled SAR images. Procedia Eng 2017; 201: 503-510. DOI: 10.1016/j.proeng.2017.09.473.

[8] Aduenko AA, Vasileisky AS, Karelov AI, Reyer IA, Rudakov KV, Strijov VV. Algorithms of detection and registration of persistent scatterers in satellite radar images. Computer Optics 2015; 39(4): 622-630. DOI: 10.18287/0134-2452-2015-39-4-622-630.

[9] Kuznetsova A, Maleva T, Soloviev V. Using YOLOv3 algorithm with pre- and post-processing for apple detection in fruit-harvesting robot. Agronomy 2020; 10: 10-16. DOI: 10.3390/agronomy10071016.

[10] Rauf HT, Saleem BA, Lali MI, Khan MA, Sharif M, Bukhari D. A citrus fruits and leaves dataset for detection and classification of citrus diseases through machine learning. Data Brief 2019; 26: 104-116.

[11] Andriyanov NA, Volkov AK, Volkov AK, Gladkikh AA, Danilov SD. Automatic x-ray image analysis for aviation security within limited computing resources. IOP Conf Ser: Mater Sci Eng 2020; 862: 1-6. DOI: 10.1088/1757-899X/862/5/052009.

[12] Taimur H, Bettayeb M, Akçay S, Khan S, Bennamoun M, Werghi N. Detecting prohibited items in X-Ray images: a contour proposal learning approach. Proc 2020 IEEE Int Conf on Image Processing (ICIP) 2020; 1: 1-6. DOI: 10.1109/ICIP40778.2020.9190711.

[13] Bogdanovich VA, Vostretsov AG. Theory of robust signal detection, discrimination and estimation [In Russian]. Moscow: "Fizmatlit" Publisher; 2004.

[14] Gruzman IS, Kirichuk VS, Kosykh VP, Peretyagin GI, Spector AA. Digital image processing in information systems: A tutorial [In Russian]. Novosibirsk: "NGTU" Publisher; 2000.
[15] Andriyanov NA, Vasiliev KK. Use autoregressions with multiple roots of the characteristic equations to image representation and filtering. CEUR Workshop Proc 2018; 2210: 273-281. DOI: 10.18287/1613-0073-2018-2210-273-281.
[16] Tikhonov VI. Optimal signal reception [In Russian]. Moscow: "Radio i Svyaz" Publisher; 1983.
[17] Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. J Phil Trans R Soc 1933; 231: 694-706.
[18] Vasiliev KK, Krasheninnikov VR. Adaptive algorithms for detecting anomalies on a sequence of multidimensional images [In Russian]. Computer Optics 1995; 14-15(1): 125-132.
[19] Denisova AYu, Myasnikov VV. Detecting anomalies in hyperspectral images [In Russian]. Computer Optics 2014; 38(2): 287-296. DOI: 10.18287/0134-2452-2014-38-2-287-296.
[20] Vasiliev KK. Detection of extended anomalies in multidimensional images [In Russian]. Vestnik UlGTU 2006; 3: 47-49.
[21] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521: 436-444. DOI: 10.1038/nature14539.
[22] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. Backpropagation applied to handwritten Zip Code recognition. Neural Computation 1989; 1(4): 541-551.
[23] Bozinovski S, Ante F. The influence of pattern similarity and transfer learning upon training of a base perceptron B2. Proceedings of Symposium Informatica 1976; 3: 121-126.
[24] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proc 26th Conf on Neural Information Processing Systems (NeurIPS) 2012; 1: 1106-1114. DOI: 10.1145/3065386.
[25] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2014; 1: 580-587.
[26] Girshick R. Fast R-CNN. Proc Int Conf on Computer Vision (ICCV) 2015; 1: 1440-1448. DOI: 10.1109/ICCV.2015.169.
[27] RenS, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. Proc 29th Conf on Neural Information Processing Systems (NeurIPS) 2015; 1: 91-99.
[28] Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proc IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2016; 1: 779-788.
[29] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, Berg A. SSD: Single shot multibox detector. Proc European Conf on Computer Vision (ECCV) 2016; 1: 1-17. DOI: 10.1007/978-3-319-46448-0_2.
[30] SSD-Mobile. Source: ⟨https://github.com/IntelAI/models/tree/master/benchmarks/object_detection/tensorflow/ssd-mobilenet⟩.
[31] Tensor Flow 2 detection zoo. Source: ⟨https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf2_detection_zoo.md⟩.
[32] Chaban LN. Methods and algorithms for pattern recognition in automated decryption of remote sensing data: a tutorial [In Russian]. Moscow: "MIIGAiK" Publisher; 2016.
[33] Zhao Z, Zheng P, Xu S, Wu X. Object detection with deep learning: A review. IEEE Trans Neural Netw Learn Syst 2019; 30(11): 3212-3232. DOI: 10.1109/TNNLS.2018.2876865.
[34] Sharma K, Nileshsingh T. A review and an approach for object detection in images. Int J Comput Vis Robot 2017; 7: 196-237. DOI: 10.1504/IJCVR.2017.10001813.
[35] Zou Z, Zhenwei S, Yuhong G, Jieping Y. Object detection in 20 years: A Survey. Source: ⟨https://arxiv.org/pdf/1905.05055.pdf⟩.
[36] Hlova M. How to find a reliable computer vision development company. Source: ⟨https://www.n-ix.com/computer-vision-development-company/⟩.
[37] Vasiliev KK. Optimal discrete time signal processing [In Russian]. Moscow: "Radiotekhnika" Publisher; 2016.
[38] Soifer VA, ed. Methods of computer image processing [In Russian]. Moscow: "Fizmatlit" Publisher; 2003. ISBN: 5-9221-0270-2.
[39] Ramchandran A, Sangaiah AK. Unsupervised anomaly detection for high dimensional data—An exploratory analysis, intelligent data-centric systems, computational intelligence for multimedia big data on the cloud with engineering applications. In Book: Sangaiah AK, Sheng M, Zhang Z, eds. Computational intelligence for multimedia big data on the cloud with engineering applications. London: Academic Press; 2018: 233-251. DOI: 10.1016/B978-0-12-813314-9.00011-6.
[40] Andriyanov NA. Analysis of the acceleration of neural networks inference on intel processors based on OpenVINO Toolkit. Proc IEEE 2020 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO) 2020; 1: 1-4. DOI: 10.1109/SYNCHROINFO49631.2020.9166067.
[41] Rosebrock A. Intersection over Union (IoU) for object detection. Source: ⟨https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/⟩.
[42] Hui J. mAP (mean Average Precision) for object detection. Source: ⟨https://jonathan-hui.medium.com/map-mean-average-precision-for-object-detection-45c121a31173⟩.
[43] Powers D. Evaluation: From precision, recall and f-measure to ROC, Informedness, markedness & correlation. Journal of Machine Learning Technologies 2011; 2(1): 37-63.
[44] Powers D. What the F-measure doesn't measure: Features, flaws, fallacies and fixes. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1503.06410⟩.
[45] Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York, NY: Oxford University Press; 2003.
[46] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Proc 23rd Int Conf on Machine Learning 2006; 1: 1-8.
[47] Hoiem D, Santosh K, Hays J. Pascal VOC 2008 Challenge. Source: ⟨http://www.wisdom.weizmann.ac.il/~/vision/courses/2010_2/papers/Hoiem_et_al_Pascal08.pdf⟩.
[48] COCO Dataset. Source: ⟨https://cocodataset.org/#detection-eval⟩.
[49] Everingham M, Van Gool L, Williams C, Winn J, Zisserman A. The PASCAL visual object classes (VOC) challenge. Int J Comput Vis 2010; 88: 303-338. DOI: 10.1007/s11263-009-0275-4.
[50] Yaroslavsky LP. Digital signal processing in optics and holography: An introduction to digital optics [In Russian]. Moscow: "Radio i Svyaz" Publisher; 1987.
[51] Akhmetshin AM, Fedorenko AE. Application of the theory of Markov random fields for segmentation of multispectral images of the Earth's surface [In Russian]. Source: ⟨http://gis.nmu.org.ua/lit/doc2.doc⟩.

[52] Bychkov AA, Ponkin VA. Image detection of spatially extended objects shading the background [In Russian]. Avtometriya 1992; 4: 33-40.
[53] Egorov VA, Bartalev SA, Burtsev MA, Efremov VYu, Lupyan EA, Mazurov AA, Matveev AM. High spatial resolution satellite image referencing streaming technology [In Russian]. Modern Problems of Remote Sensing of the Earth from Space 2010; 7(4): 97-103.
[54] Luchkov NV. Development and research of algorithms for detecting extended anomalies in multispectral images [In Russian]. The thesis for the Candidate's degree in Technical Sciences; 2012.
[55] Soyfer VA. Advanced information technologies for Earth remote sensing [In Russian]. Samara: "Novaya Technica" Publusher; 2015.
[56] Bouman CA. Model based imaging processing. Purdue University Publisher; 2013.
[57] Pyatkin VP, Sapov GI. Nonparametric statistical approach to the problem of detecting some structures in aerospace images [In Russian]. Science-Intensive Technologies 2002; 3: 52-58.
[58] Vasiliev KK, Dementyev VE. Algorithms for optimal detection of signals with unknown levels in multispectral images [In Russian]. Digital Signal Processing and its Applications 2006; 2: 433-436.
[59] Brokshtein IM, Merzlyakov SN, Popova NR. Detection and localization of small objects against a non-uniform background [In Russian]. Digital Optics. Image and Field Processing in Experimental Research 1996; 3: 67-72.
[60] Lei Z, Liu JC, Chan AK, Smith W. Object-based image segmentation using DWT/RDWT multiresolution Markov random field. IEEE Int Conf on Acoustics, Speech, and Signal Processing 1999; 6: 3485-3488. DOI: 10.1109/ICASSP.1999.757593.
[61] Ma WY, Manjunath BS. Edge Flow: A framework of boundary detection and image segmentation. Proc IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1997; 1: 744-749. DOI: 10.1109/CVPR.1997.609409.
[62] Zlobin VK, Eremeev VV. Vasiliev VM. Stochastic satellite imagery model and its use for segmentation of natural objects [In Russian]. Avtometriya 2001; 2: 13-15.
[63] Korolev EE, Kochergin AM, Kuznetsov AE. Automatic segmentation of cloud objects on images of the earth's surface with high spatial resolution [In Russian]. Modern Problems of Science and Education 2014; 5. Source: ⟨https://science-education.ru/pdf/2014/5/604.pdf⟩.
[64] Andriyanov NA, Vasiliev KK, Dementiev VE. Anomalies detection on spatially inhomogeneous polyzonal images. CEUR Workshop Proc 2017; 1901: 10-15. DOI: 10.18287/1613-0073-2017-1901-10-15.
[65] Andriyanov NA, Dementiev VE. Developing and studying the algorithm for segmentation of simple images using detectors based on doubly stochastic random fields. Pattern Recognition and Image Analysis 2019; 29(1): 1-9. DOI: 10.1134/S105466181901005X.
[66] Horton M, Cameron-James M, Williams R. Multiple classifier object detection with confidence. Proc 20th Australian Joint Conference on Artificial Intelligence 2007; 1: 559-568.
[67] Kirichuk VS, Parfenyuk SV, Angerov VYu. Detection of small-sized objects by sequences of TV-Images of the IR range [In Russian]. Proc 5th Int Scientific and Technical Conf on Pattern Recognition and Scene Analysis 2002; 1: 273-278.
[68] Gonzalez R, Woods R. Digital image processing [In Russian]. Moscow: "Tekhnosphera" Publisher; 2012.
[69] Vasyukov VN, Gruzman IS, Rayfeld MA, Spektor AA. New approaches to solving problems of image processing and recognition [In Russian]. Science-Intensive Technologies 2002; 3: 44-51.
[70] Andriyanov NA, Dementiev VE, Vasiliev KK. Developing a filtering algorithm for doubly stochastic images based on models with multiple roots of characteristic equations. Pattern Recognition and Image Analysis 2019; 29(1): 10-20. DOI: 10.1134/S1054661819010048.
[71] Vasil'ev KK, Dement'ev VE, Andriyanov NA. Application of mixed models for solving the problem on restoring and estimating image parameters. Pattern Recognition and Image Analysis 2016; 26(1): 240-247. DOI: 10.1134/S1054661816010284.
[72] Andriyanov NA, Dementyiev VE. Determination of borders between objects on satellite images using a two-proof doubly stochastic filtration. J Phys: Conf Ser 2019; 1353: 1-6. DOI: 10.1088/1742-6596/1353/1/012006.
[73] Vasil'ev KK, Dement'ev VE, Andriyanov NA. Doubly stochastic models of images. Pattern Recognition and Image Analysis 2015; 25(1): 105-110. DOI: 10.1134/S1054661815010204.
[74] Vasiliev KK, Dementiev VE, Andriyanov NA. Filtration and restoration of satellite images using doubly stochastic random fields. CEUR Workshop Proc 2016; 1814: 10-20.
[75] Shcherbakov MA, Panov AA. Nonlinear filtering with adaptation to local image properties. Computer Optics 2014; 38(4): 818-824. DOI: 10.18287/0134-2452-2014-38-4-818-824.
[76] Vasilyev KK, Ageev SA. The adaptive decorrelation algorithm of signal detection. Proc 1st Int Conf on Digital Signal Processing and Its Applications 1998; 2: 133-136.
[77] Anikin IV, Shagiakhmetov MR. Methods of fuzzy processing, recognition and analysis of objects [In Russian]. Pattern Recognition and Scene Analysis: Proc 5th Int Scientific and Technical Conf 2002; 1: 16-20.
[78] Buriak DYu, Vizilter YuV. Automated design of nearly optimal procedures for identifying and detecting objects in the image using genetic algorithms [In Russian]. Proc 12th Int Conf on CG and MV Graphicon 2002; 1: 17-20.
[79] Buriak DYu, Vizilter YuV Decision procedure representation models and their use in a genetic algorithm for finding optimal image analysis procedures [In Russian] Methods and Means of Information Processing: Proc 1st All-Russian Scientific Conf 2003; 1: 317-323.
[80] Dubes RC, Jain AK, Nadabar SG, Chen CC. MRF model-based algorithms for image segmentation. Proc 10th Int Conf on Pattern Recognition (ICPR) 1990; 1: 808-814. DOI: 10.1109/ICPR.1990.118221.
[81] Yahne B. Digital image processing [In Russian]. Moscow: "Tekhnosphera" Publisher; 2007.
[82] Beggel L, Pfeiffer M, Bischl B. Robust anomaly detection in images using adversarial autoencoders. Source: ⟨https://ecmlpkdd2019.org/downloads/paper/581.pdf⟩.
[83] Ghoneim S. React token-based authentication module with Axios Interceptors. Source: ⟨https://medium.com/@salma_ghoneim⟩.
[84] Reed IS, Yu X. Adaptive multiple-band CFAR detection of an optical pattern with unknown spectral distribution. IEEE Trans Signal Process 1990; 38(10): 1760-1770.
[85] Basener W, Ientilucci E, Messinger DW. Anomaly detection using topology. Proc SPIE 2007; 6565:16-32. DOI: 10.1117/12.745429.
[86] Schaum AP. Hyperspectral anomaly detection beyond RX. Proc SPIE 2007; 6565:122-130. DOI: 10.1117/12.718789.
[87] Zheltov SY, Vizilter YuV, Ososkov MV, Beketova IV, Karateev SL. Automatic selection of human face and its characteristic features in color digital images [In Russian]. Bulletin of Computer and Information Technologies 2005; 10: 2-7.

[88] Belim SV, Kutlunin PE. Boundary extraction in images using a clustering algorithm [In Russian]. Computer Optics 2015; 39(1): 119-124. DOI: 10.18287/0134-2452-2015-39-1-119-124.
[89] Shapiro L, Stockman J. Computer vision = Computer Vision [In Russian]. Moscow: "Binom. Laboratoria Znanii" Publisher; 2006.
[90] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Conf on Computer Vision and Pattern Recognition 2001; 1:1-9. DOI: 10.1.1.10.6807.
[91] Padilla R, Filho C, Costa M. Evaluation of Haar cascade classifiers for face detection. Int Conf on Digital Image Processing (ICDIP) 2012; 6: 466-469.
[92] Lingua A, Marenchino D, Nex F. Performance analysis of the SIFT operator for automatic feature extraction and matching in photogrammetric applications. Sensors 2009; 9(5): 3745-3766. DOI: 10.3390/s90503745.
[93] Qu X, Soheilian B, Habets E, Paparoditis N. Evaluation of SIFT and SURF for vision based localization. ISPRS Int. Arch. Photogramm. Remote Sens Spat Inf Sci 2016; XLI-B3: 685-692. DOI: 10.5194/isprs-archives-XLI-B3-685-2016.
[94] Calonder M, Lepetit V, Strecha C, Fua P. BRIEF: Binary robust independent elementary features. In Book: Daniilidis K, Maragos P, Paragios N, eds. Computer Vision – ECCV 2010. Berlin, Heidelberg: Springer-Verlag; 2010: 778-792. DOI: 10.1007/978-3-642-15561-1_56.
[95] Khoi P, Thien LH, Viet VH. Face retrieval based on local binary pattern and its variants: A comprehensive study. Int J Adv Comput Sci Appl 2016; 7: 249-258. DOI: 10.14569/IJACSA.2016.070632.
[96] Karaaba M, Surinta O, Schomaker L, Wiering MA. Robust face recognition by computing distances from multiple histograms of oriented gradients. Proc 2015 IEEE Symposium Series on Computational Intelligence 2015; 1: 203-209.
[97] Face recognition: from traditional to deep learning methods [In Russian]. Source: ⟨https://russianblogs.com/article/1856282938/⟩.
[98] Felzenszwalb P, Girshick R, McAllester D, Ramanan D. Object detection with discriminatively trained part based models. TPAMI 2010; 32(9): 1627-1645. DOI: 10.1109/TPAMI.2009.167.
[99] Fischler M, Elshlager R. The representation and matching of pictorial structures. IEEE Trans on Computer 1973; 22(1): 67-92. DOI: 10.1109/T-C.1973.223602.
[100] Parkhi OM, Vedaldi A, Zisserman A, Jawahar C. Cats and dogs. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) 2012; 1: 3498-3505. DOI: 10.1109/CVPR.2012.6248092.
[101] He K, Gkioxari G, Dollár P, Girshick R. MaskR-CNN. Source: ⟨https://arxiv.org/abs/1703.06870⟩.
[102] Chervyakov NI, Lyakhov PA, Nagornov NN, Valueva MV, Valuev GV. Hardware implementation of a convolutional neural network using computations in the residue number system. Computer Optics 2019; 43(5): 857-868. DOI: 10.18287/2412-6179-2019-43-5-857-868.
[103] Zhang Z. Derivation of backpropagation in convolutional neural network (CNN). Source: ⟨https://pdfs.semanticscholar.org/5d79/11c93ddcb34cac088d99bd0cae9124e5dcd1.pdf⟩.
[104] Dwarampudi M, Subba NV. Reddy effects of padding on LSTMs and CNNs. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1903.07288⟩.
[105] Christlein V, Spranger L, Seuret M, Nicolaou A, Král P, Maier A. Deep generalized max pooling. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1908.05040⟩.
[106] Selective search. Source: ⟨https://www.koen.me/research/selectivesearch/⟩.
[107] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2016; 1: 1-12. DOI: 10.1109/CVPR.2016.90.
[108] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press; 2000.
[109] Simonyan K. Zisserman A. Deep convolutional networks for large-scale image recognition. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1409.1556⟩.
[110] Girshick R. Fast R-CNN. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1504.08083⟩.
[111] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1506.01497⟩.
[112] Mask R-CNN: modern neural network architecture for object segmentation in images [In Russian]. Source: ⟨https://habr.com/ru/post/421299/⟩.
[113] Single shot detectors review. Source: ⟨https://towardsdatascience.com/review-ssd-single-shot-detector-object-detection-851a94607d11⟩.
[114] Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1804.02767⟩.
[115] Bochkovskiy A, Wang C, Liao HM. YOLOv4: Optimal speed and accuracy of object detection. arXiv Preprint. Source: ⟨https://arxiv.org/abs/2004.10934⟩.
[116] YOLOv5 Object detection. Source: ⟨https://laptrinhx.com/guide-to-yolov5-for-real-time-object-detection-142707357/⟩.
[117] YOLOv5 release. Source: ⟨https://github.com/ultralytics/yolov5⟩.
[118] BCCD. Source: ⟨https://public.roboflow.com/object-detection/bccd⟩.
[119] Tan M, Pang R, Le QV. EfficientDet: Scalable and efficient object detection. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1911.09070⟩.
[120] Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1905.11946⟩.
[121] Law H, Deng J. CornerNet: Detecting objects as paired keypoint. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1808.01244⟩.
[122] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1708.02002⟩.
[123] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1612.03144⟩.
[124] Bochkovskiy A, Wang C, Liao HY. YOLOv4: Optimal speed and accuracy of object detection. arXiv Preprint. Source: ⟨https://arxiv.org/pdf/2004.10934v1.pdf⟩.
[125] Object detection on COCO test-dev. Source: ⟨https://paperswithcode.com/sota/object-detection-on-coco⟩.
[126] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Zh, Lin S, Guo B. Swin transformer: Hierarchical vision transformer using shifted windows. arXiv Preprint. Source: ⟨https://arxiv.org/pdf/2103.14030v1.pdf⟩.
[127] Zhou X, Koltun V, Krahenbuhl P. Probabilistic two-stage detection. arXiv Preprint. Source: ⟨https://arxiv.org/pdf/2103.07461v1.pdf⟩.

[128] Wang C, Bochkovskiy A, Liao H. Scaled-YOLOv4: Scaling cross stage partial network. arXiv Preprint. Source: ⟨https://arxiv.org/pdf/2011.08036v2.pdf⟩.
[129] Liu Y, Wang S, Liang T, Zhao Q, Tang Z, Ling H. CBNet: A novel composite backbone network architecture for object detection. arXiv Preprint. Source: ⟨https://arxiv.org/pdf/1909.03625v1.pdf⟩.
[130] Du X, Lin T, Jin P, Ghiasi G, Tan M, Cui Y, Le QV, Song X. SpineNet: Learning scale-permuted backbone for recognition and localization. Proc IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2020; 1: 11593-11601. DOI: 10.1109/CVPR42600.2020.01161.
[131] Wang J, Sun K, Cheng T, Jiang B, Deng C, Zhao Y, Liu D, Mu Y, Tan M, Wang X, Liu W, Xiao B. Deep high-resolution representation learning for visual recognition. IEEE Trans Pattern Anal Mach Intell 2020; 1: 1-23. DOI: 10.1109/tpami.2020.2983686.
[132] Fang H, Sun J, Wang R, Gou M, Li Y, Lu C, Tong SJ. InstaBoost: Boosting instance segmentation via probability map guided copy-pasting. Proc 2019 IEEE/CVF Int Conf on Computer Vision (ICCV) 2019; 1: 682-691. DOI: 10.1109/ICCV.2019.00077.
[133] Gao Z, Wang L, Wu G. LIP: Local importance-based pooling. Proc 2019 IEEE/CVF Int Conf on Computer Vision (ICCV) 2019; 1: 3355-3364. DOI: 10.1109/ICCV.2019.00345.
[134] Vu T, Jang H, Pham T, Yoo C. Cascade RPN: Delving into high-quality region proposal network with adaptive convolution. Proc 33rd Conf on Neural Information Processing Systems (NeurIPS 2019) 2019; 1: 1-11.
[135] Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. Proc 2017 IEEE Conf on Computer Vision and Pattern Recognition (CVPR) 2017; 1: 7263-7271. DOI: 10.1109/CVPR.2017.690.
[136] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. Int Conf on Learning Representations 2021; 1: 1-22.
[137] Thuan D. Evolutoin of YOLO algorithm and YOLOv5: the state of the art object detection. Source: ⟨https://www.theseus.fi/bitstream/handle/10024/452552/Do_Thuan.pdf⟩.
[138] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv Preprint. Source: ⟨https://arxiv.org/abs/1810.04805⟩.
[139] Brown T, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language models are few-shot learners. arXiv Preprint. Source: ⟨https://arxiv.org/pdf/2005.14165.pdf⟩.
[140] Beal J, Kim E, Tzeng E, Park DH, Zhai A, Kislyuk D. Toward transformer-based object detection toward transformer-based object detection. arXiv Preprint. Source: ⟨https://arxiv.org/abs/2012.09958⟩.
[141] Gan Z, Chen Y, Li L, Zhu C, Cheng Y, Liu J. Large-scale adversarial training for vision-and-language representation learning: Supplementary material. Proc 34th Conf on Neural Information Processing Systems (NeurIPS) 2020; 1: 1-5.
[142] Mehta P. Multimodal deep learning fusion of multiple modalities using deep learning. Source: ⟨https://towardsdatascience.com/multimodal-deep-learning-ce7d1d994f4⟩.
[143] Ray A, Rajeswar S, Chaudhury S. Text recognition using deep BLSTM networks. Proc Eighth Int Conf on Advances in Pattern Recognition (ICAPR) 2015; 1: 1-6. DOI: 10.1109/ICAPR.2015.7050699.
[144] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997; 9: 1735-80. DOI: 10.1162/neco.1997.9.8.1735.
[145] Bianco S, Cadene R, Celona L, Napoletano P. Benchmark analysis of representative deep neural network architectures. IEEE Access 2018; 4: 1-8. DOI: 10.1109/ACCESS.2018.2877890.

Сведения об авторах

Андриянов Никита Андреевич, 1990 года рождения, доцент департамента анализа данных и машинного обучения Финансового университета при Правительстве Российской Федерации. В 2013 году окончил с отличием Ульяновский государственный технический университет по специальности «Инфокоммуникационные технологии и системы связи». В 2017 году защитил диссертацию на соискание ученой степени кандидата технических наук. Область научных интересов: обработка изображений, интеллектуальный анализ данных, статистическая обработка сигналов. E-mail: *nikita-and-nov@mail.ru* .

Дементьев Виталий Евгеньевич, 1982 года рождения, зав. каф. телекоммуникации Ульяновского государственного технического университета. В 2007 году защитил диссертацию на соискание ученой степени кандидата технических наук, в 2020 году – диссертацию на соискание ученой степени доктора технических наук. Сфера научных интересов: статистический анализ изображений, распознавание образов. E-mail: *dve@ulntc.ru* .

Ташлинский Александр Григорьевич, 1954 года рождения, зав. каф. радиотехники Ульяновского государственного технического университета. В 1984 году защитил диссертацию на соискание степени кандидата технических наук, в 1999 году – диссертацию на соискание степени доктора технических наук. Сфера научных интересов: адаптивные стохастические процедуры оценивания параметров изображений и сигналов, статистический анализ изображений, распознавание образов. E-mail: *tag@ulstu.ru* .

ГРНТИ: 28.23.15
Поступила в редакцию 13 мая 2021 г. Окончательный вариант – 7 августа 2021 г.

Detection of objects in the images: from likelihood relationships towards scalable and efficient neural networks

N.A. Andriyanov [1], V.E. Dementiev [2], A.G. Tashlinskiy [2]
[1] Financial University under the Government of the Russian Federation,
125993, Moscow, Russia, Leningradskiy pr-t 49;
[2] Ulyanovsk State Technical University,
432027, Ulyanovsk, Russia, Severny Venets 32

Abstract

The relevance of the tasks of detecting and recognizing objects in images and their sequences has only increased over the years. Over the past few decades, a huge number of approaches and methods for detecting both anomalies, that is, image areas whose characteristics differ from the predicted ones, and objects of interest, about the properties of which there is a priori information, up to the library of standards, have been proposed. In this work, an attempt is made to systematically analyze trends in the development of approaches and detection methods, reasons behind these developments, as well as metrics designed to assess the quality and reliability of object detection. Detection techniques based on mathematical models of images are considered. At the same time, special attention is paid to the approaches based on models of random fields and likelihood ratios. The development of convolutional neural networks intended for solving the recognition problems is analyzed, including a number of pre-trained architectures that provide high efficiency in solving this problem. Rather than using mathematical models, such architectures are trained using libraries of real images. Among the characteristics of the detection quality assessment, probabilities of errors of the first and second kind, precision and recall of detection, intersection by union, and interpolated average precision are considered. The paper also presents typical tests that are used to compare various neural network algorithms.

Keywords: pattern recognition, object detection, computer vision, image processing, random fields, CNN, IoU, mAP, probability of correct detection.

Citation: Andriyanov NA, Dementiev VE, Tashlinskii AG. Detection of objects in the images: from likelihood relationships towards scalable and efficient neural networks. Computer Optics 2022; 46(1): 139-159. DOI: 10.18287/2412-6179-CO-922.

Acknowledgements: This study was partly funded by the Russian Foundation of Basic Research under projects ## 20-17-50020 and 19-29-09048.

Authors' information

Nikita Andreevich Andriyanov (b. 1990), Associate Professor of Data Analysis and Machine Learning department at the Financial University under the Government of the Russian Federation. In 2013, he graduated with honors from Ulyanovsk State Technical University with a degree in Information and Communication Technologies and Communication Systems. In 2017 he defended his thesis for the degree of Candidate of Technical Sciences. Research interests: image processing, data mining, statistical signal processing. E-mail: *nikita-and-nov@mail.ru* .

Vitaly Evgenievich Dementiev (b.1982), head of the Telecommunication department at Ulyanovsk State Technical University. In 2007 he defended his thesis for the degree of Candidate of Technical Sciences, in 2020 – the thesis for the degree of Doctor of Technical Sciences. Research interests: statistical analysis of images, pattern recognition.
E-mail: *dve@ulntc.ru* .

Alexandr Grigorievich Tashlinskiy (b. 1954), head of the Radio Engineering department at Ulyanovsk State Technical University. In 1984 he defended his thesis for the degree of Candidate of Technical Sciences, in 1999 – the thesis for the degree of Doctor of Technical Sciences. Research interests: adaptive stochastic procedures for evaluating the parameters of images and signals, statistical analysis of images, pattern recognition. E-mail: *tag@ulstu.ru* .

Received May 13, 2021.The final version – August 7, 2021.