DOI 10.37882/2223–2966.2021.04.29

ОБНАРУЖЕНИЯ ОБЪЕКТОВ НА ОСНОВЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧЕ НАВИГАЦИИ АВТОНОМНОГО МОБИЛЬНОГО РОБОТА

Найнг Мин Тун
Аспирант, Московский государственный технический университет имени Н.Э. Баумана, г. Москва
naingminhtun52@gmail.com
Гаврилов Александр Игоревич
К.т.н, доцент, Московский государственный технический университет имени Н.Э. Баумана, г. Москва
alexgavrilov@mail.ru
Пья Пьо Паинг
Аспирант, Московский государственный технический университет имени Н.Э. Баумана, г. Москва
ppaing12@gmail.com
Ньян Линн Тун
Аспирант, Московский государственный технический университет имени Н.Э. Баумана, г. Москва
nyanlin54@gmail.com
Тхет Аунг Тху
Аспирант, Московский государственный технический университет имени Н.Э. Баумана, г. Москва
thetathu@gmail.com

OBJECT DETECTION FOR AN AUTONOMOUS MOBILE ROBOT BASED ON DEEP NEURAL NETWORKS

Naing Min Tun
A. Gavrilov
Pyae Phyo Paing
Nyan Linn Tun
Thet Aung Thu

Summary. Object detection is the most important visual task for an autonomous mobile robot. Computer vision, which includes images and videos, can serve as a cheaper sensor for detecting objects than others. In this paper, we used a small object dataset and conducted end-to-end own multiclass object detection using deep neural networks. We used a pre-trained VG 16 model using two new fully-connected subnets, not only for feature extraction, but also for the calculations of bounding box and object classification. Finally, the effectiveness of the proposed model was evaluated on test images using the accuracy measurement metric (mAP).

Keywords: object detection, convolutional neural networks, deep learning.

Аннотация. Обнаружение объектов является важнейшей задачей для автономного мобильного робота. Компьютерное зрение, включающее в себя изображения и видео, может служить более дешевым датчиком для обнаружения объектов, чем другие. В данной работе использован небольшой набор данных объектов и проведено мульти-классовое обнаружение объектов с помощью глубоких нейронных сетей. Предварительная обученная модель VGG16 дополненная двумя полно-связанными подсетями используется не только для выделения признаков, но и для вычисления ограничивающего прямоугольника и классификация объектов. Проведены оценки эффективности предложенной модели на тестовых изображениях с использованием метрики измерения точности (mAP).

Ключевые слова: обнаружение объектов, свёрточные нейронные сети, глубокое обучение.

Введение

Безопасная и надежная автономная система вождения для мобильного работа основана на точном восприятии окружающей среды. Автономное транспортное средство должно точно обнаруживать автомобили, пешеходов, препятствия, дорожные знаки и другие объекты в режиме реального времени, чтобы принимать правильные управленческие решения, обеспечивающие безопасность. Обнаружение объектов является важнейшей задачей для автономного вождения. Различные решения автономных транспортных средств могут иметь различные комбинации датчиков восприятия, но обнаружение объектов на основе изображений практически незаменимо. Датчики изображения стоят недорого по сравнению с другими, такими как LiDAR. Данные изображений (включая видео) гораздо более многочисленны, чем, например, LiDAR облачные точки, и их гораздо легче собирать и аннотировать. Таким образом, датчик обнаружения объектов на основе изображения является хорошим выбором для автономного мобильного робота.

К теме обнаружения объектов, в связи с тесной связью с видеоанализом и пониманием изображений, было привлечено большое внимание исследователей в последние годы. Для получения полного понимания изображения, необходимо сосредоточиться не только на классификации различных изображений, но и попытаться точно оценить расположение объектов, содержащихся в каждом изображении. Эта задача называется обнаружением объектов [1], которое обычно состоит из различных подзадач, таких как обнаружение лиц [2], обнаружение пешеходов [3] и обнаружение скелетов [4]. Как одна из фундаментальных проблем компьютерного зрения, обнаружение объектов способно предоставить ценную информацию для семантического понимания изображений и видео, что связано со многими приложениями, включая классификацию изображений [5], [6], анализ поведения человека [7], распознавание лиц [8] и т.д.

В данной работе построена модель обнаружения объектов для автономного мобильного робота на основе компьютерного зрения с использованием нейронных сетей.

Методы обнаружения объектов

Традиционные методы обнаружения объектов основаны на ручном детектировании признаков и неглубоких обучаемых архитектурах. Их производительность легко стагнирует при построении сложных ансамблей, объединяющих множество низкоуровневых признаков изображения с высокоуровневым контекстом от детекторов объектов и классификаторов. С быстрым развитием глубокого обучения для решения проблем, существующих в традиционных архитектурах, вводятся более мощные инструменты, которые способны изучать семантические, высокоуровневые, более глубокие признаки. Эти модели ведут себя по-разному в сетевой архитектуре, в стратегии обучения, функции оптимизации и т.д.

Однако из-за больших различий в ракурсах наблюдения, позах, окклюзиях и условиях освещения трудно идеально выполнить обнаружение объекта с помощью дополнительной задачи локализации объекта. К решению именно этой задачи было привлечено внимание исследователей в последние годы [9–12]. Задача обнаружения объектов состоит в том, чтобы определить, где находятся объекты на заданном изображении (локализация объектов) и к какой категории относится каждый объект (классификация объектов). Таким образом, традиционные модели обнаружения объектов можно в основном разделить на три этапа: выбор информативной области, выделение признаков и классификация.

Выбор информативной области. Поскольку различные объекты могут появляться в любых положениях на изображении и иметь различные пропорции или размеры, естественным выбором является сканирование всего изображения с помощью многомасштабного скользящего окна. Хотя эта исчерпывающая стратегия позволяет выяснить все возможные положения объектов, ее недостатки также очевидны. Большое количества окон-кандидатов может привести к существенным вычислительным затратам и обработке избыточных окон. Однако если применяется только фиксированное количество шаблонов скользящих окон, могут быть получены неудовлетворительные регионы.

Выделение признаков. Чтобы распознать различные объекты, нужно выделить визуальные признаки, которые могут обеспечить семантическое и надежное представление. Репрезентативными являются методы SIFT [13], HOG [14] и признаки Хаара [15]. Однако из-за разнообразия внешнего вида, условий освещения и фона трудно вручную разработать надежный дескриптор признаков, чтобы идеально описать все виды объектов.

Классификация. Классификатор необходим для того, чтобы выделить целевой объект из всех других категорий и сделать представления более иерархичными, семантическими и информативными для визуального распознавания. Обычно хорошим выбором являются метод опорных векторов (SVM) [16], AdaBoost [17] и модель на основе деформируемых деталей (DPM) [18].

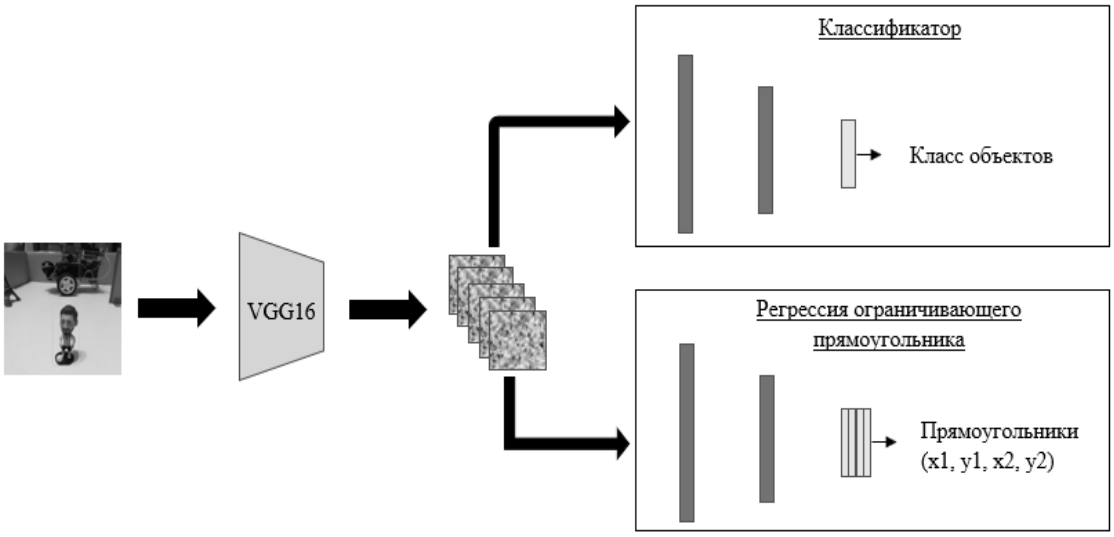

Рис. 1. Процесс функционирования предлагаемой модели обнаружения объектов.

Благодаря использованию глубоких нейронных сетей (deep neural networks, DNNs) [6][7], значительный выигрыш получается при введении регионов с признаками CNN (R-CNN) [9]. DNN, или наиболее репрезентативные CNN, действуют иначе, чем традиционные подходы. Они обладают более глубокими архитектурами, способными моделировать изучать сложные функции. Кроме того, робастные алгоритмы обучения позволяют изучать информативные представления объектов без необходимости проектировать объекты вручную [19].

Был предложен ряд улучшенных моделей R-CNN, включая архитектуру Fast R-CNN, который совместно оптимизирует задачи классификации и регрессии ограничивающего прямоугольника [10], Faster R-CNN использует дополнительную подсеть для генерации предложений регионов [12] и YOLO, который выполняет обнаружение объектов с помощью регрессии с фиксированной сетью [11]. Все эти модели обеспечивают различные степени улучшения производительности обнаружения по сравнению с базовым R-CNN и делают более достижимым точное обнаружение объектов в реальном времени.

Предложенная в данной работе модель основана на предварительно обученной модели VGG16, отличающейся наличием двух полно-связанных подсетей, предназначенных для классификации объектов и прогнозирования локализации объектов.

Решение задачи обнаружения объектов

Предварительно обученная модель VGG16 была использована для обучения собственного много-классового детектора объектов с использованием регрессии ограничивающего прямоугольника. Для создания архитектуры заданной модели необходимо модифицировать базовую сетевую архитектуру предварительно обученной модели VGG16 следующим образом:

- Принять предварительно обученную модели VGG16 от ImageNet и удалить полно-связанные слои.
- Разместить две новых полно-связанных подсети на выходе основной модели VGG16. Одну для ограничивающего прямоугольника ((x, y) координаты), а другую для прогнозирования меток классов.
- Провести тонкую настройку (fine-tune) всей сети для сквозного обнаружения объектов.

На рис. 1. показан процесс функционирования предлагаемой модели на основе предварительно обученной модели VGG16.

Первоначально, чтобы обучить надежный классификатор, нужно много изображений, отличающихся друг от друга. Они должны иметь разные фоны, содержать случайные объекты и различные условия освещения. Было сделано около 50 снимков для каждого отдельного объекта. Около 80 процентов изображений для целей обучения, а остальные 20 процентов для целей тестирования. Для того чтобы создать метку и ограничивающую рамку обучающих данных, было использовано программное обеспечение для маркировки изображений под названием LabelImg.

Все обучающие изображения пропускаются через ряд слоев предварительно обученной модели VGG16 (т.е. свёрточный, пулинг) для извлечения значимых

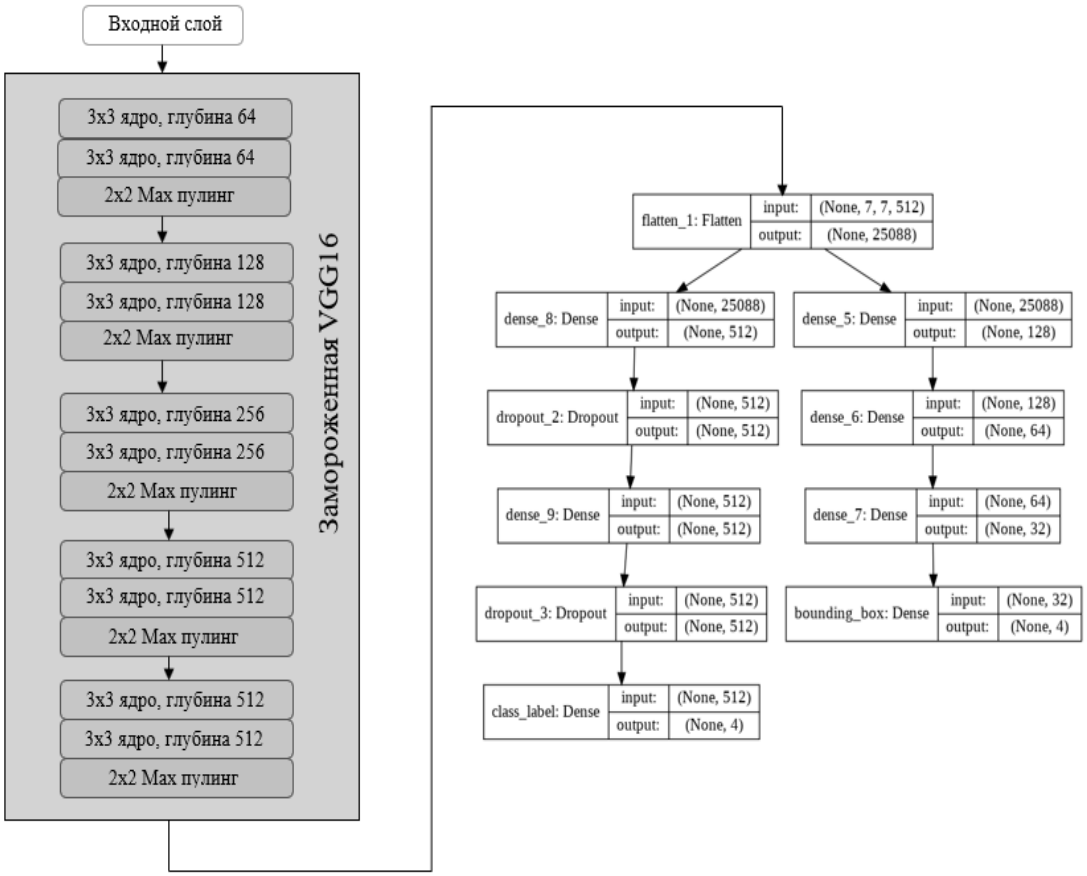

Рис. 2. Сетевая архитектура мульти-классового обнаружения объектов.

признаков, которые были пропущены через два различных полно-связанных слоя. Для классификатора была использована функция потери softmax. Функция softmax нормализует входные данные и генерирует распределение вероятностей для выходов. Операция softmax может быть представлена следующим образом:

$$softmax(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Где, $\vec{z}$ — входной вектор, e^{z_i} — стандартная экспоненциальная функция для входного вектора, e^{z_j} — стандартная экспоненциальная функция для выходного вектора, K — количество классов мульти-классового классификатора.

Функция потери softmax называется кросс-энтропийной потерей, поскольку ее функция потери используется для получения потери по вероятностям. Функция потери выглядит следующим образом:

$$L = J(\hat{y}, y) = \sum_{i=0}^{K} y_i log\hat{y}_i = -\sum_{i=0}^{K} y_i log(softmax(\vec{z})_i)$$

Где, $\hat{y}$ — прогнозируемое значение, y — желаемое значение. Для получения градиента этой потери softmax для обновления нейронных сетей была использована производная функция, следующим образом:

$$\frac{\partial L}{\partial z} = \hat{y} - y$$

Для регрессии ограничивающего прямоугольника была использована функция потери L2, используемая для минимизации ошибки, которая представляет собой сумму всех квадратов разностей между желаемым значением и прогнозируемым значением:

$$L = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Создана мульти-классовая модель обнаружения объектов на основе предварительно обученной модели VGG16 с использованием Keras API, которая включает в себя два новые полно-связанные подсети для классификатора и регрессии ограничивающего прямоугольника. Подсеть для ограничивающего прямоугольника состоит из 128, 64, 32 и 4 узлов соответственно.

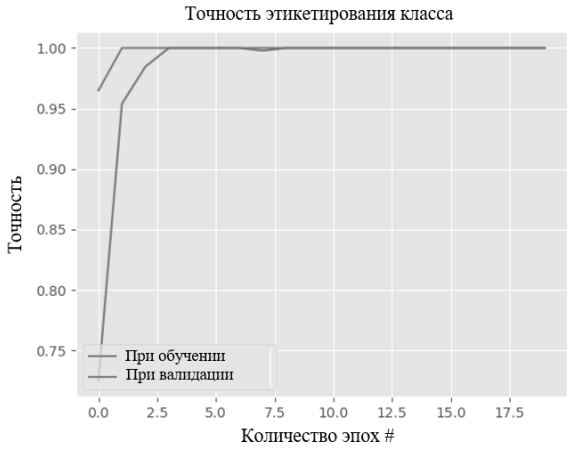

Рис. 3. точность этикетирования класса.

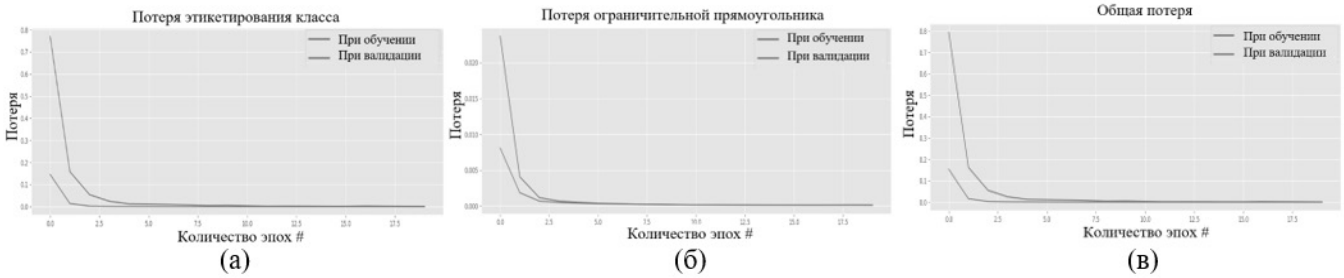

Рис. 4. Три компонента потерь. а) потеря этикетирования класса, б) потеря ограничивающего прямоугольника и в) общая потеря (представляет собой комбинацию потери этикетирования класса и ограничивающего прямоугольника)

Наиболее важной частью подсети регрессии ограничивающего прямоугольника является конечный слой, который содержит 4 нейрона, соответствующие координатам (x, y) для верхнего левого и нижнего правого углов ограничивающего прямоугольника. За прогнозирование классовой метки обнаруженного объекта отвечала подсеть классификатора. Эта подсеть состоит из 512, 512 и 4 узлов соответственно. В последнем слое были установлены также 4 нейрона из-за четырех различных категорий обучающих данных в нашей работе. Сетевая архитектура предлагаемой модели представлена на рис. 2.

Включен ряд графиков для визуализации выходных данных процесса обучения предлагаемой модели. На рис. 3 показана точность этикетирования класса и на рис. 4 представлены результаты потери (потеря этикетирования класса, потеря ограничительной рамки прямоугольника и общая потеря).

Здесь можно отметить, что предлагаемая модель правильно классифицирует метки обнаруженных объектов в обучающем и тестовом наборе со 100% точностью.

Для проверки предложенной модели обнаружения объекта были использованы некоторые изображения тестового подмножества собственного набора данных, включенного в категории мини-образцов с целью применения в проекте управления движением транспортного средства. После тестирования предложенной модели на тестовых изображениях каждой категории были получены различные результаты, включая пересечение над союзом (Intersection over union — IoU), точность (precision), полнота (recall), средняя средняя точность (mean average precision — mAP), как показано в таблице 1.

Предложенная модель обнаруживает объекты широкого диапазона масштабов и пропорций. Каждый выходной прямоугольник связан с меткой категории. Время выполнения для получения этих результатов составляет 982 мс на изображение при использовании CPU процессора. На рис. 5 показаны результаты про-

Таблица 1. Сравнение результатов предлагаемой модели обнаружения объектов в соответствии с категориями тестовых изображений

Категория (образец)	Количество тестовых изображений	IoU	точность	полнота	mAP
Человек	39	62.28	91.36	29.26	43.73
Блок	23	70.07	96.49	39.90	45.06
Дерево	33	91.37	100.00	31.44	54.54
Башня	16	86.86	100.00	32.22	54.54
Всего	111	76.09	96.02	30.00	44.33

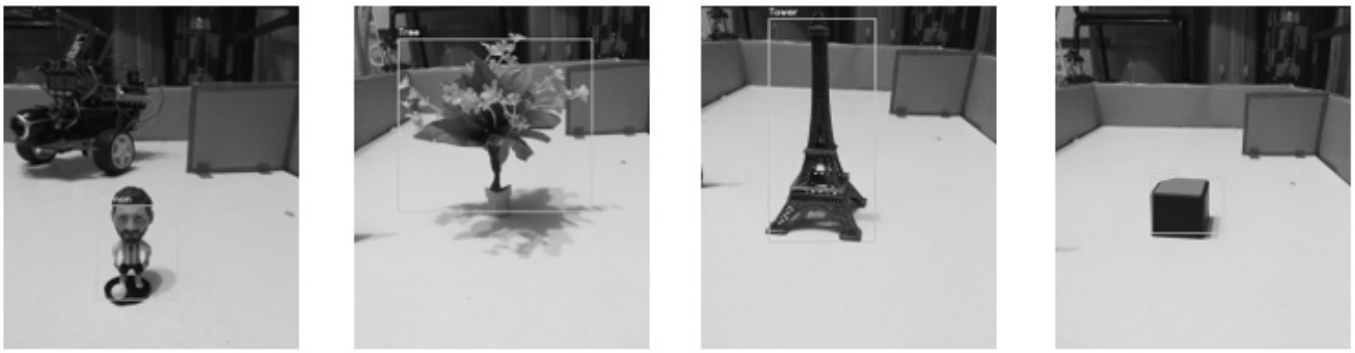

Рис. 5. Примеры результатов обнаружения объектов на собственном наборе данных с использованием предложенной модели.

верки на тестовом подмножестве собственного набора данных.

Согласно полученным результатам, предложенная модель может обнаруживать объекты с надежными результатами, связанными с меткой класса и ограничивающим прямоугольником. Для более точного определения ограничивающего прямоугольника необходимо усовершенствовать модель. Тем не менее, модель подходит для применения в проектах управления автономным транспортным средством и в образовательных целях.

Заключение

В работе представлена система мульти-классового распознавания объектов с использованием VGG16 модели, у которой устранены полно-связанные слоя и добавлены две полно-связанные подсети для прогнозирования меток классов и генерации ограничивающего прямоугольника. Построена структура предлагаемой модели для сквозного обнаружении объектов с помощью библиотеки Keras. По результатам эксперимента можно заметить, что предложенная модель может обеспечить надежные результаты классификации и приемлемые результаты регрессии ограничивающего прямоугольника даже при небольшом объеме обучающих данных.

Основным преимуществом предлагаемой модели является возможность применения собственных данных для мульти-классового обнаружения объектов даже с использованием маломощных вычислительных машин. Основным недостатком предлагаемой модели является необходимость более точного формирования координат ограничивающего прямоугольника.

В работе рассмотрена способность сквозного обнаружения объектов на основе одного базовой предварительно обученной модели с двумя новыми полно-связанными подсетями. В дальнейшей работе возможности предлагаемой модели будут использоваться в системах управления автономным транспортным средством в качестве системы машинного зрения с использованием большого количества обучающихся данных.

ЛИТЕРАТУРА

1. Felzenszwalb P.F., Girshick R.B., McAllester D., Ramanan D. Object detection with discriminatively trained part-based models // IEEE transactions on pattern analysis and machine intelligence. Vol. 32. № 9. 2009. P. 1627–45.
2. Sung K.K., Poggio T. Example-based learning for view-based human face detection // IEEE Transactions on pattern analysis and machine intelligence. Vol. 20. № 1. 1998. P. 39–51.
3. Dollar P., Wojek C., Schiele B., Perona P. Pedestrian detection: An evaluation of the state of the art // IEEE transactions on pattern analysis and machine intelligence. Vol. 34. № 4. 2011. P. 743–61.
4. Kobatake H., Yoshinaga Y. Detection of spicules on mammogram based on skeleton analysis // IEEE Transactions on Medical Imaging. Vol. 15. № 3. 1996. P. 235–45.
5. Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S., Darrell T. Caffe: Convolutional architecture for fast feature embedding // InProceedings of the 22nd ACM international conference on Multimedia. 2014. P. 675–678.
6. Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. Vol. 25. 2012. P. 1097–105.
7. Cao Z., Simon T., Wei S.E., Sheikh Y. Realtime multi-person 2d pose estimation using part affinity fields // InProceedings of the IEEE conference on computer vision and pattern recognition. 2017. P. 7291–7299.
8. Yang Z., Nevatia R. A multi-scale cascade fully convolutional network face detector // In2016 23rd International Conference on Pattern Recognition (ICPR). IEEE. 2016. P. 633–638.
9. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation // InProceedings of the IEEE conference on computer vision and pattern recognition. 2014. P. 580–587.
10. Girshick R. Fast r-cnn // InProceedings of the IEEE international conference on computer vision. 2015. P. 1440–1448.
11. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection // InProceedings of the IEEE conference on computer vision and pattern recognition. 2016. P. 779–788.
12. Ren S., He K., Girshick R., Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks // arXiv preprint arXiv:1506.01497. 2015.
13. Lowe D.G. Distinctive image features from scale-invariant keypoints // International journal of computer vision. Vol. 60. № 2. 2004. P. 91–110.
14. Dalal N., Triggs B. Histograms of oriented gradients for human detection // In2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE. Vol. 1. 2005. P. 886–893.
15. Lienhart R., Maydt J. An extended set of haar-like features for rapid object detection // InProceedings. international conference on image processing. IEEE. Vol. 1. 2002. P. I–I.
16. Cortes C., Vapnik V. Support vector machine // Machine learning. Vol. 20. № 3.1995. P. 273–97.
17. Freund Y., Schapire R.E. A decision-theoretic generalization of on-line learning and an application to boosting // Journal of computer and system sciences. Vol. 55. № 1. 1997. P. 119–39.
18. Felzenszwalb P.F., Girshick R.B., McAllester D., Ramanan D. Object detection with discriminatively trained part-based models // IEEE transactions on pattern analysis and machine intelligence. Vol. 32. № 9. 2009. P. 1627–1645.
19. LeCun Y., Bengio Y., Hinton G. Deep learning // nature. Vol. 521. № 7553. 2015. P. 436–44.