Разработка нейросетевого алгоритма распознавания надписей на изображениях реальных сцен

В.А. Лобанова [1], Ю.А. Иванова [1]
[1] Национальный исследовательский Томский политехнический университет, 634050, Россия, г. Томск, пр. Ленина, д. 30

Аннотация

Работа посвящена проектированию и реализации нейросетевого алгоритма детектирования надписей на изображениях реальных сцен. Проведен обзор существующих нейросетевых и классических моделей, в качестве базовой была выбрана модель U-net. На ее основе предложен и реализован алгоритм детектирования текстовых областей на изображениях. В ходе проведения экспериментов были определены следующие параметры нейронной сети: размеры входных изображений, количество и типы составляющих её слоёв. В качестве предобработки рассматривались билатеральные фильтры сглаживания и сглаживающие частотные фильтры. Увеличение исходной базы изображений KAIST Scene Text Database достигается за счёт применения поворотов, сжатия и разбиения входящих в неё изображений. Полученные результаты превосходят другие методы по значению F-меры и достигают 0,88.

Ключевые слова: детектирование текстовых областей, U-Net, сегментация изображений, изображения реальных сцен.

Цитирование: **Лобанова, В.А.** Разработка нейросетевого алгоритма распознавания надписей на изображениях реальных сцен / В.А. Лобанова, Ю.А. Иванова // Компьютерная оптика. – 2022. – Т. 46, № 5. – С. 790-800. – DOI: 10.18287/2412-6179-CO-1047.

Citation: Lobanova VA, Ivanova YA. Development of software for the segmentation of text areas in real-scene images. Computer Optics 2022; 46(5): 790-800. DOI: 10.18287/2412-6179-CO-1047.

Введение

В настоящее время существует огромное количество информации, хранящейся в виде изображений, содержание которых представляет собой определенную ценность. Детектирование и последующее распознавание текста на изображениях может быть применено в таких областях, как перевод фотографий документов в текстовую форму [1], автоматическое определение номерных знаков автомобилей [2], геолокация объекта по названиям улиц, улучшение качества детектирования и распознавания объектов на изображениях. Однако объемы информации, хранящейся в виде изображений, велики, что делает невозможным ее обработку вручную. Тем не менее автоматизированные методы обработки изображений позволяют успешно справляться с этой задачей.

Несмотря на широкую область возможного применения и прогресс в сфере машинного обучения, обнаружение и сегментация текстовых областей на изображениях все еще представляет собой проблему [1 – 11]. Текст, расположенный на изображениях реальных сцен, может быть различным по размеру, стилю, цвету, повороту относительно горизонта. Также возможно его перекрытие другими объектами на изображении или низкая контрастность с фоном. Применение фильтров для сглаживания шумов и увеличения контрастности на границах объектов не является универсальным решением проблемы, так как может как вызывать ложные срабатывания, так и мешать обнаружению значимых областей.

Целью данной работы является разработка нейросетевого алгоритма распознавания надписей на изображениях реальных сцен.

1. Методы распознавания текстовых областей

Методы распознавания текстовых областей используются для определения наличия и выделения местоположения текстовых областей на изображениях. Однако точность данного определения может быть снижена из-за различных размеров, стилей и направлений надписей. Кроме того, низкий контраст с фоном или наличие сложного фона могут вызывать дополнительные затруднения. Все существующие методы распознавания текстовых областей можно разделить на следующие группы:

- методы связных компонент [4, 5, 19 – 23];
- текстурные методы [6, 8, 10, 12, 17, 22 – 26];
- методы глубокого обучения [13 – 16].

1.1. Методы связных компонент

Принцип работы методов связных компонент заключается в поиске и объединении малых компонентов в большие на основании определённых характеристик пикселей: яркость, цвет, толщина контура элемента. Затем из полученных компонентов извлекаются признаки для дальнейшей классификации на текстовые и нетекстовые компоненты. Найденные текстовые компоненты извлекаются из изображений и объединяются в текстовые области.

Главными преимуществами методов связных компонент являются простота вычислений и высокая

точность. Однако данные методы плохо работают при изменении поворота или масштабировании изображений, обработке изображений со сложным фоном.

Для рассмотрения были выбраны метод максимально стабильных экстремальных областей (MSERs) [3, 4, 19 – 21] и метод преобразования по толщине штриха символа текста (SWT) [5, 19 – 23].

Метод MSERs использует экстремальные области для распознавания текстовых областей. Экстремальные области на изображении определяются двумя условиями относительно множества пикселей внутри области [3]:

- область остаётся постоянной при преобразовании координат (поворот, растяжение);
- область остаётся постоянной при изменении яркости изображения.

Данные методы требуют малого количества памяти для реализации, могут работать в режиме реального времени, но плохо распознают текстовые области на размытых или неконтрастных изображениях, часто дают ложноположительные результаты [4].

Метод SWT строится на предположении о том, что буквы и символы на текстовых областях имеют определенные геометрические особенности: одинаковая ширина для каждого символа и одинаковая толщина штриха символа текста. Другими словами, в большинстве случаев текстовые области имеют небольшое изменение толщины штриха от символа к символу, в отличие от нетекстовых областей [5].

Метод SWT преобразует изображение в массив, где каждый элемент содержит значение толщины штриха для соответствующего пикселя, используя для этого оператор Кэнни. Далее пиксели группируются в предполагаемые текстовые области. Два соседних пикселя объединяются, если они обладают равными значениями толщины штриха. Отделить текстовые области от нетекстовых позволяет одинаковое значение толщины для текста. Затем из выделенных текстовых областей формируется предсказанное текстовое поле.

Эксперименты показывают, что метод SWT эффективен для обнаружения текста. Данный метод может быть применён к различным шрифтам и языкам, а также к различным размерам и поворотам надписей. Однако метод SWT не является полностью автоматическим и требует предварительного подбора параметров, что может привести к ложным срабатываниям для сложных случаев.

1.2. Текстурные методы

Текстурные методы основаны на идее о том, что текстовые области отличаются от фона изображений высокочастотными и регулярными текстурными признаками [6, 8, 10, 12, 17, 22 – 26]. В данных методах происходит применение классификатора к отдельным областям изображения для определения наличия или отсутствия текста. Трудности для выделения текстовых областей на изображениях могут быть вызваны тем, что наравне с тестовыми областями на изображениях реальных сцен регулярностью обладают такие объекты, как кирпичная кладка, окна домов, листва деревьев.

Для рассмотрения были выбраны следующие текстурные методы:

- дискретное косинусное преобразование [6,7];
- гистограмма направленных градиентов [8];
- признаки Хаара [12].

Дискретное косинусное преобразование (ДКП) было создано для сжатия изображений. ДКП позволяет трансформировать пространство изображения в пространство свойств с более низкой размерностью [6].

Однако данный метод можно использовать и для извлечения текстовых областей из изображений. Преобразование применяется для каждого отдельного фрагмента изображения независимо друг от друга, после чего происходит объединение значений. Значения полученных частотных коэффициентов отражают локальную периодичность в области изображения. Неявная периодичность означает, что на границах возникают разрывы.

Алгоритм, предложенный *Zhong*, использует дискретное косинусное преобразование и состоит из двух основных этапов [7]:

- обнаружение предполагаемых текстовых областей в сжатой области частотных коэффициентов;
- постобработка областей.

Предложенный алгоритм обладает высокой скоростью за счёт работы с пространством свойств с более низкой размерностью и постобработкой изображений с пониженным разрешением. Однако данный алгоритм не предоставляет высокой точности при разделении областей на текстовые и нетекстовые.

Ещё одним рассматриваемым текстурным методом является использование HOG-дескрипторов или применение метода гистограмм направленных градиентов (Histogram of Oriented Gradients). Сначала происходит разделение изображения на ячейки, затем для каждого пикселя внутри ячейки рассчитывается гистограмма направлений градиентов. Объединение полученных значений для каждой ячейки называется HOG-дескриптором. Полученное градиентное изображение выделяет контуры на изображении и исключает из рассмотрения несущественную информацию [8].

Czarnek [9] в своей работе рассматривает отдельные цифры, а не полноценный текст. Такие данные легко визуализировать и интерпретировать в отличие от изображений реальных сцен, на которых, помимо текста, присутствуют и другие объекты. Поэтому метод гистограмм направленных градиентов больше подходит для предварительной обработки изображений, но не для самостоятельного применения.

Признаки Хаара используют вейвлет-преобразование Хаара. Данный метод использует в

качестве признаков прямоугольные области, разделенные на контрастные части. Данные признаки применяются путём наложения на фрагменты рассматриваемого изображения, при наложении происходит вычисление суммы интенсивностей пикселей для каждой области признака. Различие между полученными значениями позволяет распределить по категориям фрагменты изображения [10].

В большинстве случаев текст представляет собой несколько хорошо контрастирующих по горизонтали или вертикали частей на изображении независимо от цвета текста или фона. Данный факт предоставляет возможность для применения признаков Хаара для обнаружения текстовых областей [11 – 12].

Главным минусом применения данного метода для обнаружения текстовых областей является зависимость результата от начального выбора набора применяемых признаков. Позиции букв в тексте и их формы могут сильно различаться для разных изображений, что может вызвать дополнительную проблему при подготовке набора признаков.

1.3. Методы глубокого обучения

Свёрточные нейронные сети (CNN) являются классом нейронных сетей. В отличие от полносвязных нейронных сетей, где каждый нейрон предыдущего слоя связан со всеми нейронами следующего слоя, в свёрточных нейронных сетях соседние нейроны одного слоя связаны с локальной областью нейронов последующего слоя, что значительно сокращает количество настраиваемых во время обучения весов.

Это делает свёрточные сети более вычислительно эффективными, позволяя использовать в качестве входных значений изображения большего размера. Свёрточные сети также могут включать подвыборочные слои, используемые для уменьшения размеров данных за счёт объединения выходов кластеров нейронов в один входной нейрон [13].

Основным преимуществом свёрточной сети является возможность работы со сложными изображениями, так как она обладает инвариантностью к поворотам, растяжениям, изменениям яркости или контрастности [27 – 30]. Такая сеть может извлекать информацию из изображения с небольшими по сравнению с полносвязными сетями вычислительными затратами. Недостатком свёрточной нейронной сети являются её выходные значения, представляемые в виде вектора значений: данная сеть может быть использована для обнаружения наличия текста на изображении, но не для определения его местоположения.

Структура полносвёрточных нейронных сетей (FCN) схожа со структурой свёрточной нейронной сети. Главное отличие заключается в выходном слое сети. Данный слой производит объединение всех карт признаков с предыдущих слоёв для создания изображения в виде сегментированной карты, каждый сегмент которой представляет собой область определенного класса. Полученная сегментированная карта может быть использована для определения положения текстовых областей на изображении [14].

При использовании полносвёрточной нейронной сети могут появляться ложные срабатывания или пропущенные символы при очень низкой контрастности, кривизне, сильном отражении света, плотном расположении строк текста или больших промежутках между символами. Еще одним недостатком данного типа сети является быстродействие.

Выходное изображение полносвёрточной нейронной сети архитектуры U-Net совпадает по размерам со входным изображением. Отличительной особенностью U-Net является ее симметричная структура. Она состоит из сужающей и расширяющей частей, которые создают U-образную структуру. Расширяющая часть состоит из слоев повышающей дискретизации, которые позволяют обрабатывать большее количество карт признаков. В данной структуре отсутствуют полносвязные слои, в результате выходное изображение сети содержит предсказания только для тех пикселей, которые соответствуют входному изображению [15].

Основным преимуществом архитектуры U-Net является наличие большего количества слоев повышающей дискретизации, которые позволяют обрабатывать большее количество карт признаков. Кроме того, U-Net обеспечивает попиксельную точность благодаря одинаковому размеру входных и выходных изображений [16].

Было проведено сравнение всех вышеописанных методов. Методы, основанные на поиске связных компонент, обладают высокой скоростью работы, однако плохо работают с различными искажениями изображений и имеют среднюю точность. Текстурные методы обладают средней скоростью работы, плохо работают с различными искажениями и имеют низкую точность. Методы глубокого обучения являются самой медленной группой из-за длительного времени обучения, но хорошо работают с различными искажениями изображений, а также обладают высокой точностью [27 – 30]. Следовательно, методы, основанные на глубоком обучении, представляют большую ценность для реализации.

Свёрточные нейронные сети предъявляют меньше требований к оборудованию по сравнению с другими сетями, но являются недостаточными для определения местоположения текстовых областей на изображениях. Полносвёрточные нейронные сети с несимметричной структурой и сети архитектуры U-Net находятся на одном уровне. Однако симметричная архитектура сети U-Net предоставляет больше возможностей для распознавания текстовых областей ввиду наличия большего количества слоев повышающей дискретизации для обработки карт признаков. По этой причине архитектура U-Net была выбрана для реализации алгоритма в данной работе.

2. Проектирование алгоритма

В качестве базы изображений была выбрана база KAIST Scene Text Database, из которой было выбрано 1215 фотографий, содержащих текст на английском и корейском языках, на которых после уменьшения размера сохранялся текст [17]. Изображения были получены с помощью цифровой камеры высокого разрешения и мобильного телефона с низким разрешением. Все изображения были приведены к размеру 384×384 пикселя.

Для определения расположения текста для каждого изображения представлено изображение-маска, на котором черным цветом выделен фон, а красным – текст (рис. 1).

Рис. 1. Пример исходных изображений обучающей выборки и изображений масок

В работе было проведено искусственное увеличение базы изображений за счёт поворотов и обрезки изображений. Повороты производились в обе стороны от минус 18° до плюс 18° с шагом в 6°. Размер базы после увеличения составляет 8505 изображений. Маски были бинаризированы.

Дополнительным способом увеличения базы изображений являлась подача на вход сети разбитых на 4 части с перекрытием исходных изображений. Размер итоговой базы составил 42525 изображений. Далее база изображений делится на обучающую, валидационную и тестовую в соотношении 72 %, 18 % и 10 % соответственно.

На исходных изображениях часто присутствуют такие высокочастотные составляющие, как кирпичная кладка стен, оконные рамы, деревья, что часто дает ложноположительные результаты при детектировании текстовых областей. В связи с этим было решено подвергнуть исходные изображения предварительной фильтрации.

Предлагаемый алгоритм состоит из следующих этапов:

- предварительная обработка базы изображений;
- детектирование текстовых областей.

В предварительную обработку входит загрузка данных, применение фильтров к входным изображениям, загрузка изображений-масок, на основе которых далее происходит обучение сети.

Для обучения используются 2 конфигурации сети: первая представлена на рис. 2 и используется для конечного обучения сети U-Net; вторая (меньшая по размеру) представлена на рис. 3 и используется для проверочных обучений.

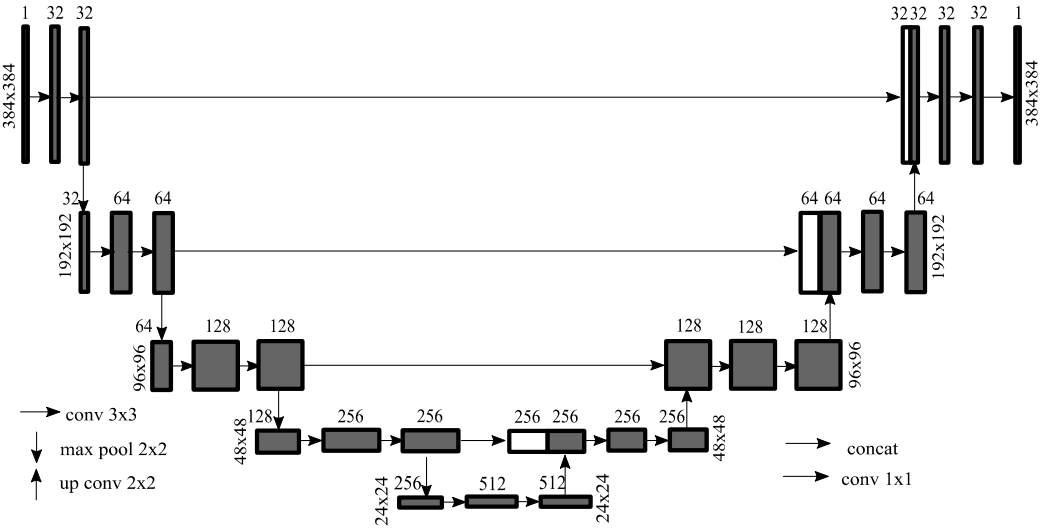

Рис. 2. Предлагаемый вариант свёрточной нейронной сети U-Net

Конфигурация сети для конечного обучения описывается далее. Входными значениями являются изображения размерами 384×384 пикселя. Первый свёрточный слой представляет собой 32 карты свёрточного слоя равного размера, получаемые с помощью ядра размером 3×3. Первый подвыборочный слой представляет собой операцию выборки локальных максимальных значений с помощью ядра размером 2×2 и шагом, соответствующим размеру ядра. Со второго по пятый свёрточный слои происходит увеличение количества карт: 64, 128, 256, 512 соответственно. Со второго по четвёртый подвыборочный слой никаких изменений не происходит.

Далее добавляются слои конкатенации карт признаков из кодирующей части. Вместо свёрточных слоёв используются слои обратной свёртки, которые

увеличивают размер изображения в 2 раз по обеим осям. С шестого по девятый слой обратной свёртки происходит уменьшение количества карт: 256, 128, 64, 32 соответственно.

В свёрточных слоях и слоях обратной свёртки используется функция активации ReLU, расчет которой производится по формуле 1:

$$f(x)=\begin{cases}0, x<0;\\ x, x\geq 0.\end{cases} \tag{1}$$

На последнем уровне свёртка используется для формирования выходного сегментированного изображения, равного по размерам входному изображению. Используется сигмоидная функция активации в форме гиперболического тангенса.

Ввиду довольно большого времени обучения конечной сети была задействована вторая структура, представленная на рис. 2, 3.

Данная сеть отличается от конечной сети размерами входных изображений, а также количеством слоёв.

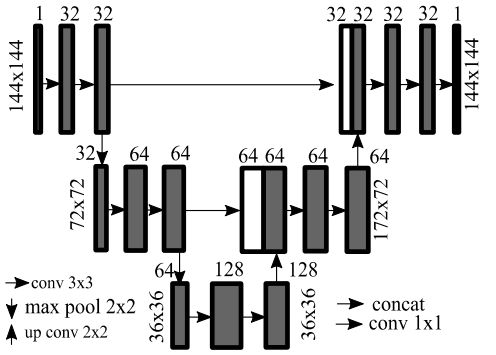

Рис. 3. Проверочная свёрточная нейронная сеть U-Net

Для вычисления ошибки сети используется коэффициент Дайса (Dice coefficient), который показывает меру сходства изображений:

$$dice=\frac{2\times\left|intersection\left(y_{true},y_{pred}\right)\right|+smooth}{\left|y_{true}\right|+\left|y_{pred}\right|+smooth}, \tag{2}$$

где y_{true} – множество истинных значений пикселей текста, y_{pred} – множество предсказанных значений пикселей текста, *intersection* – пересечение множеств, $|y_{true}|$ – количество элементов множества y_{true}, *smooth* – коэффициент сглаживания.

Чем выше значение коэффициента Дайса, тем большее количество истинных и предсказанных значений пикселей изображений совпадает, соответственно, лучше детектируются текстовые области.

В качестве дополнительных параметров оценивания работы сети было решено ввести следующие значения:

$$Presicion=\frac{TP}{TP+FP}, \tag{3}$$

где *Precision* – точность, *TP* – истинно положительное решение (определенный как текстовая область пиксель действительно является текстовым), *FP* – ложно положительное решение (определенный как текстовая область пиксель действительно не является текстовым).

$$Recall=\frac{TP}{TP+FN}, \tag{4}$$

где *Recall* – полнота, *TP* – истинно положительное решение, *FN* – ложно отрицательное решение (определенный как нетекстовая область пиксель действительно является текстовым).

$$F=\left(\beta^2+1\right)\frac{Presicion\times Recall}{\beta^2\times Presicion+Recall}, \tag{5}$$

где *F* – F-мера, β – коэффициент от 0 до ∞.

3. Настройка параметров модели

Экспериментальное обучение сети проводилось по двум направлениям: формирование конечной структуры сети, определение размера и типов предобработки входных изображений. Формирование структуры сети заключалось в определении размеров входных изображений, количества и типов слоёв. Для предобработки изображений использовались фильтры сглаживания на основе свёртки, сглаживающие частотные фильтры, разбиение и сжатие изображений.

3.1. Определение размеров входных изображений

Определение размеров входных изображений производилось на исходной базе изображений KAIST Scene Text Database [17]. Для свёрточных слоёв размер ядра составляет 3×3 пикселя, а с помощью подвыборочных слоёв производится уменьшение изображения в 2 раза. Сеть с данной архитектурой применялась для определения размеров входных изображений. Сравнительное обучение производилось для сети с 10 слоями, поэтому размеры входных изображений должны были без остатка делиться как на 3, так и на 24. Были выбраны следующие размеры входных изображений:192×192 пикселя и 384×384 пикселя.

На рис. 4 представлены графики зависимости коэффициента Дайса на обучающей и валидационной выборках от количества эпох. Для изображений размером 192×192 пикселей было проведено обучение в 110 эпох, для изображений размером 384×384 пикселей – 60 эпох.

Дальнейшее обучение не проводилось ввиду того, что среднее изменение точности для последних 10 эпох не превышало 0,06 % и 0,02 % соответственно. Следовательно, такие значения несоизмеримы с затраченным временем. Видно, что размеры изображений 384×384 пикселя показывают лучшие результаты как при равном количестве эпох, так и при равном времени обучения. Таким образом, было принято решение в дальнейшем подавать на вход сети изображения размером 384×384 пикселя.

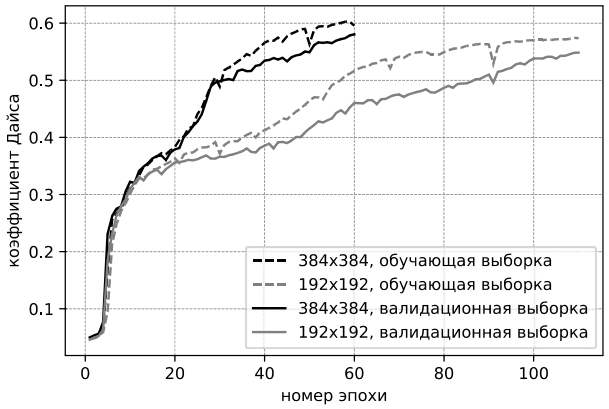

Рис. 4. Сравнительное обучение конечной сети для различных размеров входных изображений

3.2. Определение наилучшей конфигурации сети

Следующим шагом формирования структуры конечной свёрточной сети был выбор количества слоёв. Было проведено сравнительное обучение для сети с 6, 8 и 10 слоями. Большее количество слоёв не бралось в рассмотрение ввиду увеличения времени обучения. Во всех представленных случаях на вход сети подавались изображения, приведённые к размерам 384×384 пикселя. Обучение производилось в течение 10 эпох. На рис. 5 представлены графики зависимости коэффициента Дайса обучения и валидации от количества эпох.

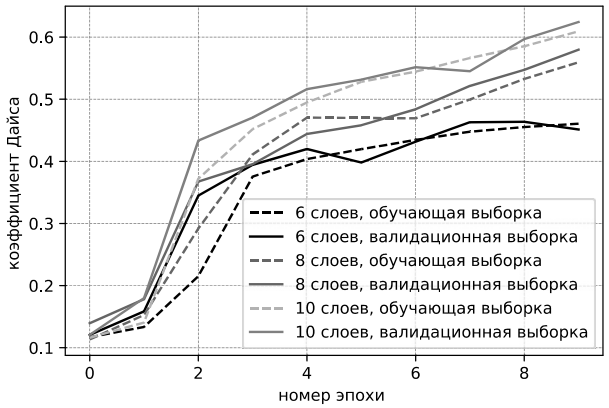

Рис. 5. Графики зависимости коэффициента Дайса обучения и валидации от количества эпох для 6, 8 и 10 слоёв

На момент последней эпохи коэффициент Дайса для обучения с 6, 8 и 10 слоями достигает 46,1 %, 55,9 % и 60,9 % соответственно. Видно, что наилучшие результаты даёт сеть с использованием 10 слоёв. Дальнейшее проверочное обучение не требуется, так как в течение последних 5 эпох обучения значения коэффициента Дайса для сети с 10 слоями в среднем превышают значения для сети с 6 и 8 слоями на 27,7 % и 12,1 % соответственно.

3.3. Определение типов слоев

Было проведено сравнение для конфигураций сети с использованием слоёв Max Pooling и без них. Во втором случае свёртка слоёв происходила за счёт слоёв Convolution2D. Обучение производилось в течение 10 эпох. На рис. 6 представлены графики зависимости коэффициента Дайса обучения и валидации от количества эпох.

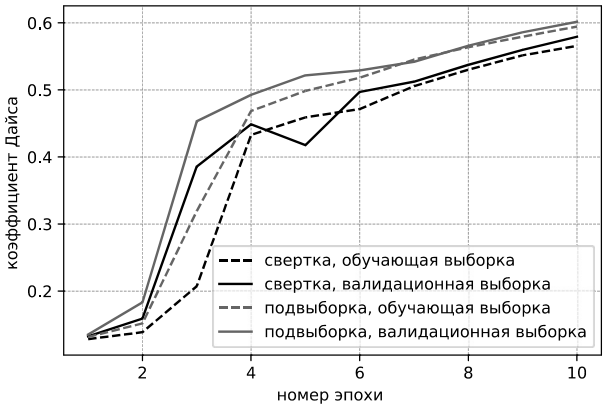

Рис. 6. Графики зависимости коэффициента Дайса обучения и валидации от количества эпох при использовании слоёв Max Pooling и без данных слоёв

На момент последней эпохи коэффициент Дайса обучения при использовании слоёв Max Pooling и без данных слоёв достигает 59,5 % и 56,6 % соответственно. Видно, что лучшие результаты обучения достигаются при использовании слоёв Max Pooling.

3.4. Проведение экспериментов по предварительной обработке изображений

Наличие на изображениях различных шумов, пониженной контрастности и других искажений может снизить точность обучаемой сети, поэтому перед подачей на сеть необходимо производить предобработку изображений.

В качестве фильтра сглаживания применялся билатеральный фильтр с настраиваемыми параметрами: $\sigma_{пр}$ – стандартное отклонение по координатному пространству, $\sigma_{яр}$ – стандартное отклонение по цветовому пространству. В качестве фильтра выделения краев применялся лапласиан (рис. 7*б*), фильтр повышения резкости представлен на рис. 7*а*.

0	-1	0
-1	5	-1
0	-1	0

(а)

0	-1	0
-1	4	-1
0	-1	0

(б)

Рис. 7. Маска фильтра повышения резкости (а), маска фильтра выделения границ Лапласа (б)

Было решено использовать следующие комбинации фильтров в качестве предобработки исходных изображений [18]:

а) билатеральный фильтр ($\sigma_{пр} = \sigma_{яр} = 50$);

б) билатеральный фильтр ($\sigma_{пр} = \sigma_{яр} = 150$);

в) билатеральный фильтр ($\sigma_{пр} = \sigma_{яр} = 50$) + фильтр резкости;

г) билатеральный фильтр ($\sigma_{пр} = \sigma_{яр} = 50$) + фильтр выделения границ;

д) билатеральный фильтр ($\sigma_{пр}=\sigma_{яр}=50$ + фильтр выделения границ).

Для сравнения предобработки входных изображений было произведено обучение на проверочной сети на исходной базе изображений длиною в 20 эпох. На рис. 8 и 9 представлены графики зависимости коэффициента Дайса от количества эпох соответственно.

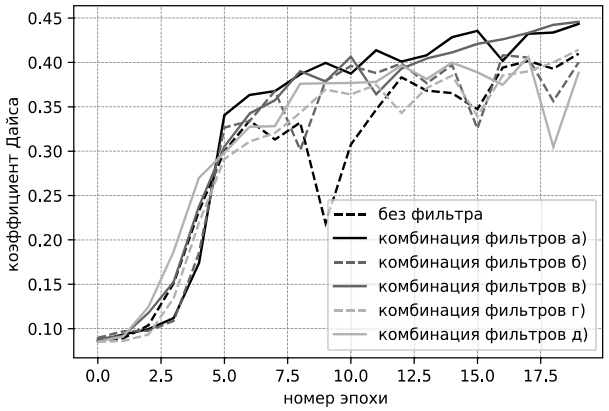

Рис. 8. Графики зависимости коэффициента Дайса обучения от количества эпох для различных фильтров

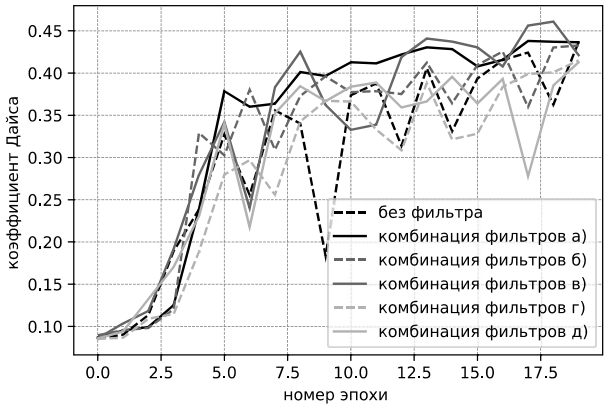

Рис. 9. Графики зависимости коэффициента Дайса валидации от количества эпох для различных фильтров

Видно, что для обучающей выборки наиболее точными и стабильными являются билатеральный фильтр сглаживания с коэффициентом 50 и с коэффициентом 150. Применение фильтров повышения резкости снижает качество детектирования.

Для валидационной выборки большинство фильтров достаточно непостоянны, самым стабильным является билатеральный фильтр сглаживания с коэффициентом 50.

Таким образом, было решено использовать билатеральный фильтр сглаживания с коэффициентом 50 для обучения конечной сети.

3.5. Сглаживающие частотные фильтры

В качестве альтернативного варианта предобработки входных изображений был рассмотрен частотный анализ с преобразованием Фурье.

Для использования преобразования Фурье как части сглаживающих фильтров были выбраны фильтры:

- идеальный фильтр низких частот;
- фильтр низких частот Гаусса;
- фильтр низких частот Баттерворта.

Их параметры: D_0 – частота среза и n – порядок – представлены в табл. 1.

Табл. 1. Фильтры размытия и их параметры

Название фильтра низких частот	Значения параметра D_0	Значения параметра n
Идеальный фильтр	80, 120, 160, 200	-
Фильтр Гаусса	120, 160, 200	-
Фильтр Баттерворта	120, 160, 200	2, 5, 8

В результате проверочных обучений были выбраны лучшие значения параметров и проведено их сравнение с фильтром сглаживания. На рис. 10 показано сравнение результатов проверочного обучения нейронной сети для лучших фильтров с преобразованием Фурье и фильтра сглаживания.

Видно, что коэффициенты Дайса даже лучших фильтров с преобразованием Фурье (фильтр Баттерворта с $D_0=160$ и $n=5$, фильтр Баттерворта с $D_0=200$ и $n=8$) не превышают значения коэффициентов для фильтра сглаживания.

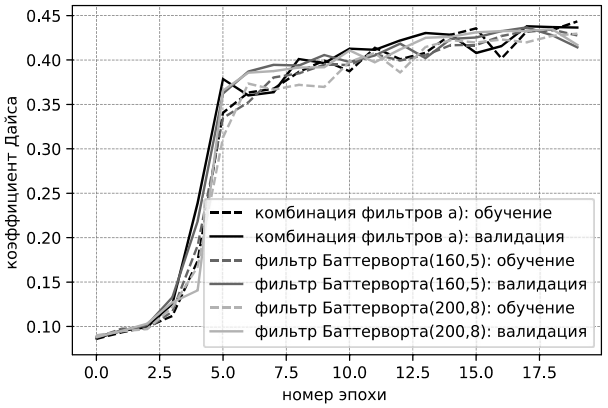

Рис. 10. Сравнение результатов проверочного обучения сети для лучших фильтров с преобразованием Фурье и фильтра сглаживания

3.6. Разбиение и сжатие изображений

Так как исходный размер изображений равен 640×480 пикселей, а на вход сети подаются изображения размером 384×384 пикселя, то было решено проводить разбиение изображений на 4 части вместо изменения их размеров перед подачей на вход сети.

Было проведено проверочное обучение длиною в 10 эпох для 3 вариантов подачи входных изображений: сжатие, разбиение, комбинация сжатия и разбиения. На рис. 11 представлены графики коэффициента Дайса для каждого варианта. Видно, что наилучший результат показывает комбинация сжатия и разбиения.

После определения параметров сети и способов предобработки изображений было проведено обучение на конечной сети длиною в 5 эпох. В качестве сравнения также было проведено обучение на конечной сети при подаче только сжатых изображений. Сравнение производилось как для равного количества

эпох обучения, так и для равного количества времени обучения. В табл. 2 представлены параметры оценки обучения, валидации и тестирования.

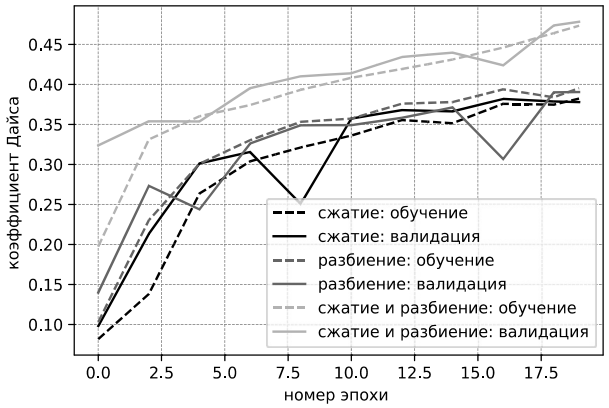

Рис. 11. Сравнение результатов проверочного обучения сети для 3 вариантов подачи входных изображений: сжатие, разбиение, комбинация сжатия и разбиения

Результаты показывают, что за одинаковое количество эпох комбинация сжатия и разбиения существенно увеличивает значения параметров оценки сети как для обучающей, так и для валидационной выборки. Сравнение с результатами предсказаний для тестовой выборки на этом этапе не может быть произведено ввиду проведения тестирования после полного завершения обучения сети.

При сравнении с результатами обучения при равном времени обучения сетей видно, что комбинация сжатия и разбиения позволяет получить лучшие результаты для коэффициента Дайса, полноты и F-меры. Точность уступает на 3,3 %, 2,9 % и 4,4 % для обучающей, валидационной и тестовой выборок соответственно.

Высокие значения полноты указывают на то, что большая часть текстовых областей действительно определяется алгоритмом; в то время как высокие значения точности указывают на то, что большая часть выделенных областей действительно являются текстовыми. Рассмотрение точности и полноты для результатов обучения при равном времени показывает, что для обучения на сжатых изображениях данные значения различаются в среднем на 12,9 %, а для обучения на комбинации сжатия и разбиения – на 2,9 %. Это означает, что при обучении только на сжатых изображениях сеть не способна выделить все текстовые области.

На рис. 12 и 13 представлены примеры лучших и худших вариантов предсказанных масок соответственно.

На большей части тестовых изображений текстовые области детектируются правильно. Сеть хорошо справляется с различными типами и размерами шрифтов, а также с поворотами. Сложными случаями для определения являются большие надписи: выделяются дополнительные области вокруг или, наоборот, опускаются области внутри символов.

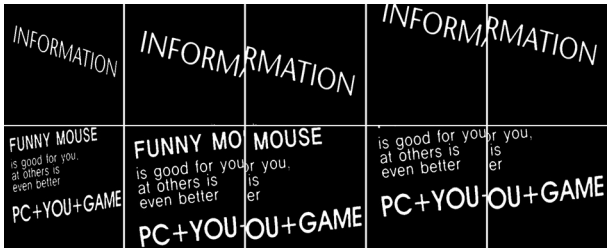

Рис. 12. Лучшие маски для тестовых изображений, полученные из сети максимальной точности

Рис. 13. Худшие маски для тестовых изображений, полученные из сети максимальной точности

Табл. 2. Сравнение обучения конечной сети

Параметр	Тип	Конечная сеть, база изображений с поворотами и сжатием		Конечная сеть, база изображений с поворотами, сжатием и разбиением
Количество эпох обучения		22	5	5
Полное время обучения, час		59	13,5	60
Коэффициент Дайса, (*smooth* = 1)	Обучение	0,897	0,605	**0,911**
	Валидация	0,897	0,660	**0,900**
	Тестирование	**0,894**	-	0,880
Точность	Обучение	**0,948**	0,584	0,915
	Валидация	**0,945**	0,813	0,917
	Тестирование	**0,946**	-	0,902
Полнота	Обучение	0,807	0,653	**0,902**
	Валидация	0,823	0,681	**0,882**
	Тестирование	0,822	-	**0,864**
F-мера (β = 1)	Обучение	0,8718	0,631	**0,908**
	Валидация	0,8798	0,662	**0,900**
	Тестирование	**0,918**	-	0,883

Было проведено сравнение результатов, полученных предложенным алгоритмом, с результатами, полученными другими авторами на базе изображений KAIST Scene Text Database [17].

Список методов, сравниваемых с предложенным алгоритмом: (1) метод, основанный на квадратичной дискриминантной функции [19]; (2) метод наименьшего разреза графа [19]; (3) комбинация метода квадратичной дискриминантной функции и наименьшего разреза графа [19]; (4) метод определения пороговых значений Otsu [24]; (5) метод определения пороговых значений Niblack [25]; (6) метод на основе кластеризации K-средних [26]; (7) метод максимально стабильных экстремальных областей [20]; (8) модифицированный метод на основе кластеризации K-средних [17]; (9) метод детектирования особых областей (Blob) [22]; (10) метод, основанный на Марковской модели случайного поля [23]; (11) комбинация метода максимально стабильных экстремальных областей и алгоритма дендрограмм одинарной кластеризации [21].

Основой методов (12 – 18) являются модели глубокого обучения, такие как генеративные состязательные сети, свёрточные нейронные сети, MaskRCNN [27 – 30].

Тестирование методов (1 – 6) производилось только на английской части базы изображений, в то время как работы (7 – 18) рассматривают полный набор изображений. Дальнейшее обращение к работам производится по названиям их методов.

В табл. 3 представлены параметры оценки тестирования для всех перечисленных методов, а также для метода, предложенного в данной работе. В качестве параметров сравнения были использованы точность, полнота и F-мера.

Табл. 3. Сравнение результатов тестирования методов на базе изображений KAIST Scene Text Database

База	Метод	Точность	Полнота	F-мера
Английская часть	(1) Метод, основанный на квадратичной дискриминантной функции [19]	0,789	**0,925**	0,851
	(2) Метод наименьшего разреза графа [19]	0,786	0,918	0,847
	(3) Комбинация метода квадратичной дискриминантной функции и наименьшего разреза графа [19]	0,828	0,893	0,860
	(4) Метод определения пороговых значений Otsu [24]	0,747	0,905	0,818
	(5) Метод определения порога Niblack [25]	0,689	0,909	0,784
	(6) Метод на основе кластеризации K-средних [26]	0,763	0,914	0,821
Полная	(7) Метод максимально стабильных экстремальных областей [20]	**0,924**	0,341	0,480
	(8) Модифицированный метод на основе кластеризации K-средних [17]	0,690	0,600	0,640
	(9) Метод детектирования особых областей (Blob) [22]	0,879	0,489	0,593
	(10) Метод, основанный на Марковской модели случайного поля [23]	0,697	0,291	0,376
	(11) Комбинация метода максимально стабильных экстремальных областей и алгоритма дендрограмм одинарной кластеризации [21]	0,670	0,890	0,760
	(12) Метод максимально стабильных экстремальных областей + преобразование ширины штриха + генеративная состязательная сеть [27]	0,85	0,84	0,84
	(13) Нелинейная нейронная сеть, основанная на свёрточной нейронной сети [28]	0,59	0,79	0,67
	(14) Генеративная состязательная сеть [29]	0,43	0,69	0,53
	(15) MaskRCNN++ [30]	0,82	0,78	0,80
	(16) MM-MaskRCNN [30]	0,76	0,84	0,70
	(17) Подход, основанный на MaskRCNN[30]	0,49	0,64	0,40
	(18) Двухпроходный детектор текста, основанный на каскаде RCNN [30]	0,78	0,82	0,74
	Предложенный метод	0,90	0,86	**0,88**

Как видно из табл. 3, максимальным значением точности обладает работа с использованием метода максимально стабильных экстремальных областей (7). Полученное значение превышает точность метода, представленного в данной работе, на 2,2 %. Однако полнота и F-мера данного метода показывают очень низкие результаты, следовательно, большое количество текстовых областей не обнаруживается. Все остальные методы показывают точность тестирования ниже, чем предложенный метод.

Все методы, тестирование которых проходит на английской части базы изображений, обладают более высокими значениями полноты. При этом разница с текущим методом не превышает 6,1 % для наилучшего результата. Это может быть обусловлено упрощенным набором данных для тестирования для данных методов. Тем не менее эти же методы показывают более низкие значения точности, что означает наличие большего числа ложных срабатываний на нетекстовых областях.

Среди методов, тестирование которых проходит на полной базе изображений, лишь метод (11) показывает значение полноты выше текущего, но значительно проигрывает по точности. Неплохие результаты показывает метод (12), хотя они несколько ниже предложенного в данной работе метода.

Отметим, что расчет метрик для методов (12) – (18) производится на основе пересечений прямоугольных областей, ограничивающих текст, а не

пересечений множества детектированных пикселей. За счет этого значение вышеупомянутых метрик будет больше даже при более низком качестве сегментации.

Предложенный метод показывает наибольшее значение F-меры. В среднем другие методы показывают значения на 6,0 % меньше.

Таким образом, метод, представленный в данной работе, не обладает наилучшими значениями точности и полноты при их раздельном рассмотрении, но при этом превосходит все остальные по значению F-меры. Это означает, что ни один метод из других работ не превосходит предложенный метод по всем параметрам.

Высокие значения сразу всех параметров показывают, что обучение происходило более сбалансировано, то есть большая часть текстовых областей действительно определяется алгоритмом и большая часть выделенных областей действительно являются текстовыми. Таким образом, предложенный алгоритм превосходит алгоритмы других авторов по F-мере, что говорит о лучшей его сбалансированности между верными и ложными срабатываниями, и является сопоставимым с лучшими результатами алгоритмов по точности и полноте.

Заключение

В ходе выполнения работы был проведен анализ предметной области, включающий в себя рассмотрение существующих методов обнаружения текстовых областей на изображениях реальных сцен. В результате в качестве классификатора была выбрана свёрточная нейронная сеть архитектуры U-Net.

В качестве базы изображений был выбран набор данных KAIST Scene Text Database, для которого было проведено увеличение количества изображений за счёт применения поворотов, сжатия и разбиения.

Был разработан нейросетевой алгоритм распознавания надписей на изображениях реальных сцен. Экспериментально были подобраны такие параметры, как размеры входных изображений, способ предварительной их обработки, конфигурация сети U-net (количество и тип составляющих её слоев).

В результате обучения предложенного нейросетевого алгоритма удалось добиться высокого значения F-меры: 0,91 для обучающей, 0,9 для валидационной и 0,88 для тестовой выборки.

Было проведено сравнение с работами других авторов, проводивших тестирование на базе изображений KAIST Scene Text Database: полный набор данных и только английская часть. В результате сравнения предложенный в данной работе алгоритм показывает наилучшие результаты по значению F-меры. Преимуществом данного алгоритма является попиксельная сегменатация изображений, что значительно упростит дальнейшее распознавание текста.

References

[1] Mechi O, Mehri M, Ingold R, Ben Amara NE. Text line segmentation in historical document images using an adaptive U–Net architecture. Int Conf on Document Analysis and Recognition 2019: 369-374.

[2] Chowdhury PN, Shivakumara P, Raghavendra R, Pal U, Lu T, Blumenstein M. A new U-Net based license plate enhancement model in night and day images 5th Asian Conf on Pattern Recognition 2019: 749-763.

[3] Matas J, Chum O, Urban M, Pajdla T. Robust wide-baseline stereo from maximally stable extremal regions. Image Vis Comput 2004; 22(10): 761-767.

[4] Neumann L, Matas J. Real-time scene text localization and recog-nition. IEEE Conf on Computer Vision and Pattern Recognition 2012: 3538-3545.

[5] Epshtein B, Ofek E, Wexler Y. Detecting text in natural scenes with stroke width transform. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2010; 2963-2970.

[6] Ahmed N, Natarajan T, RaoKR. Discrete cosine transform. IEEE Trans Comput 1974; C-23(1): 90-93.

[7] Zhong Y, Zhang H, Jain AK. Automatic caption localization in compressed video. IEEE Trans Pattern Anal Mach Intell 2000; 22(4): 385-392.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection. IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2005; 1: 886-893.

[9] Czarnek N. Physically motivated feature development for machine learning applications. Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Electrical and Computer Engineering in the Graduate School of Duke University 2017.

[10] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. Proc 2001 IEEE Computer Society Conf on Computer Vision and Pattern Recognition 2001; 1: 511-518.

[11] Ghorbel A. Generalized Haar-like filters for document analysis: application to word spotting and text extraction from comics. Document and Text Processing. Université de La Rochelle; 2016.

[12] Chen X, Yuille AL. Detecting and reading text in natural scenes. Proc 2004 IEEE Computer Society Conf on Computer Vision and Pattern Recognition 2004; 2: 366-373.

[13] Goodfellow IJ, Bulatov Y, Ibarz J, Arnoud S, Shet V. Multi-digit number recognition from street view imagery using deep convolutional neural networks. Proc Int Conf on Learning Representations 2014: 1-12.

[14] Zhang Z, Zhang C, Shen W, Yao C, Liu W, Bai X. Multi-oriented text detection with fully convolutional networks. Proc 2016 IEEE Conf on Computer Vision and Pattern Recognition 2016: 4159-4167.

[15] Ronneberger O, Fischer P, BroxT. U-net: Convolutional networks for bio-medical image segmentation. Med Image Comput Comput Assist Interv 2015; 9351: 234-241.

[16] Bezmaternykh PV, Ilin DA, Nikolaev DP. U-Net-bin: hacking the document image binarization contest. Computer Optics 2019; 43(5): 825-832. DOI: 10.18287/2412-6179-2019-43-5-825-832.

[17] Lee S, Cho MS, Jung K, Kim JH. Scene text extraction with edge constraint and text collinearity. 20th Int Conf on Pattern Recognition 2010: 3983-3986.

[18] Tomasi C, Manduchi R. Bilateral filtering for gray and color images. 6th Int Conf on Computer Vision 1998: 839-846.

[19] Bai B, Yin F, Liu CL. A seed-based segmentation method for scene text extraction. 11th IAPR Int Workshop on Document Analysis Systems2014: 262-266.
[20] Agrawal A, Mukherjee P, Srivastava S, Lall B. Enhanced characterness for text detection in the wild. Proc 2nd Int Conf on Computer Vision & Image Processing 2018: 359-369.
[21] Gomez L, Karatzas D. A fast hierarchical method for multi-script and arbitrary oriented scene text extraction. Int J Doc Anal Recognit 2016; 19(4): 335-349.
[22] Jahangiri, M., Petrou, M. An attention model for extracting components that merit identification. 2009 16th IEEE Int Conf on Image Processing (ICIP) 2009: 965-968.
[23] Li Y, et al. Characterness. An indicator of text in the wild. IEEE Trans Image Process 2014; 23(4): 1666-1677.
[24] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern Syst 1979; 9(1): 62-66.
[25] Niblack W. An introduction to digital image processing. New York: Prentice Hall; 1986.
[26] Kita K, Wakahara T. Binarization of color characters in scene images using k-means clustering and support vector machines. 2010 20th Int Conf on Pattern Recognition 2010: 3183-3186.
[27] Saha S, Chakraborty N, Kundu S, Paul S, Mollah AF, Basu S, Sarkar R. Multi-lingual scene text detection and language identification. Pattern Recognit Lett 2020; 138: 16-22.
[28] Li L, Yu S, Zhong L, Li X. Multilingual text detection with nonlinear neural network. Math Probl Eng 2015; 2015: 431608.
[29] Xu H, Su X, Liu T, Guo P, Gao G, Bao F. A natural scene text extraction approach based on generative adversarial learning. Int Conf on Neural Information Processing 2019: 65-73.
[30] Nayef N, Patel Y, Busta M, Chowdhury PN, Karatzas D, Khlif W, Matas J, Pal U, Burie J-C, Liu C-l, Ogier JM. ICDAR2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. 2019 Int Conf on Document Analysis and Recognition (ICDAR) 2019: 1582-1587.

Сведения об авторах

Лобанова Виктория Александровна, 1997 года рождения, в 2021 году окончила Национальный исследовательский Томский политехнический университет по направлению 09.04.01 «Информатика и вычислительная техника». Область научных интересов: обработка графических изображений, нейронные сети, программирование. E-mail: *val17@tpu.ru* .

Иванова Юлия Александровна, 1986 года рождения, в 2009 году окончила Томский политехнический университет по специальности «Программное обеспечение вычислительной техники и автоматизированных систем», к.т.н., доцент отделения информационных технологий. Область научных интересов: обработка и анализ изображений, нейросетевые модели, методы машинного обучения. E-mail: *jbolotova@tpu.ru* .

ГРНТИ: 28.23.37.
Поступила в редакцию 13 сентября 2021 г. Окончательный вариант – 22 апреля 2022 г.

Development of software for the segmentation of text areas in real-scene images

V.A. Lobanova [1], Yu.A. Ivanova [1]
[1] Tomsk Polytechnic University, 634050, Tomsk, Russia

Abstract

This article discusses the design and development of a neural network algorithm for the segmentation of text areas in real-scene images. After reviewing the available neural network models, the U-net model was chosen as a basis. Then an algorithm for detecting text areas in real-scene images was proposed and implemented. The experimental training of the network allows one to define the neural network parameters such as the size of input images and the number and types of the network layers. Bilateral and low-pass filters were considered as a preprocessing stage. The number of images in the KAIST Scene Text Database was increased by applying rotations, compression, and splitting of the images. The results obtained were found to surpass competing methods in terms of the F-measure value.

Keywords: deep learning, U-Net architecture, image processing, image segmentation, text areas, real scenes images.

Citation: Lobanova VA, Ivanova YA. Development of software for the segmentation of text areas in real-scene images. Computer Optics 2022; 46(5): 790-800. DOI: 10.18287/2412-6179-CO-1047.

Authors' information

Viktoriya Aleksandrovna Lobanova (b. 1997) graduated from Tomsk Polytechnic University in 2021, majoring in Informatics and Computer Engineering. Research interests are computer graphics processing, neural networks, programming. E-mail: *val17@tpu.ru* .

Yuliya Aleksandrovna Ivanova, (b. 1986), graduated from Tomsk Polytechnic University in 2009, majoring in Informatics and Computer Science. Works as associate professor in Tomsk Polytechnic University. Research interests: image analysis and processing, neural networks, machine learning algorithms. E-mail: *jbolotova@tpu.ru* .

Received September 13, 2021. The final version – April 22, 2022.