Информационные процессы, Том 25, № 4, 2025, стр. 860–873.

ТЕОРИЯ И МЕТОДЫ ОБРАБОТКИ ИНФОРМАЦИИ

Компактная нейросетевая модель для детекции текстовых строк на изображениях документов на основе быстрого преобразования Хафа

А. В. Гайер[*,**], **А. В. Шешкус**[*,**], **Д. П. Николаев**[*,**], **В. В. Арлазаров**[*,**]

[*] *Федеральный исследовательский центр "Информатика и управление" Российской академии наук, г. Москва, Россия*

[**] *ООО "Смарт Энджинс Сервис" , г. Москва, Россия*

Поступила в редколлегию 26.11.2025 г. Принята 15.12.2025 г.

Аннотация—Современные системы распознавания документов основаны на нейросетевых моделях, размер и вычислительная сложность которых затрудняют их применение на широком спектре устройств. В данной работе предлагается сверхкомпактная нейросетевая модель для детекции текста в средах с ограниченными ресурсами. В ее основе лежит обработка глобальных признаков с линейной структурой, соответствующих прямым текстовым строкам, в пространстве Хафа. Размер модели составляет всего 116 килобайт — что в 8 раз меньше, чем детектор текста MULDT, и в 41 раз меньше, чем детектор текста в PaddleOCR. Проведенные эксперименты на наборах данных FUNSD, SROIE, SVRD и XFUND показывают, что предложенная модель имеет сопоставимое качество с современными компактными детекторами текста.

КЛЮЧЕВЫЕ СЛОВА: Распознавание документов, глубокое обучение, быстрое преобразование Хафа (БПХ), детекция текста.

DOI: 10.53921/18195822_2025_25_4_860

1. ВВЕДЕНИЕ

Распознавание и анализ документов широко применяются в финансовых, логистических, торговых и прочих организациях [1,2]. При этом содержащаяся в документах конфиденциальная и персональная информация накладывает ограничения на способы распознавания: передача изображения на сервер небезопасна и может привести к утечке данных. Поэтому необходимо гарантировать работу систем распознавания на мобильных устройствах, с помощью которых делается фотография — включая малопроизводительные смартфоны и планшеты. Помимо аппаратных ограничений в виде вычислительной мощности устройств, существуют также программные ограничения в некоторых средах: например, в WebAssembly (WASM). Данная технология широко применяется в веб-версиях банковских приложений для запуска вычислительно сложных систем, написанных на высокопроизводительных языках программирования. Такие модули запускаются локально на устройстве прямо в среде браузера. Однако производительность WASM приложений ниже в сравнении с нативными, особенно в случае параллельных вычислений. Также для веб-приложений крайне критичен размер программ: WASM-модули загружаются на устройство в момент обращения к ресурсу, и поэтому большие нейросетевые модели в составе системы распознавания значительно увеличивают время загрузки приложения и ухудшают пользовательский опыт. Таким образом, потребность в быстрых и компактных моделях чрезвычайно высока.

В данной работе рассматривается один из ключевых этапов любой системы распознавания и анализа документов — детектирование текста. Эта задача является вычислительно сложной,

поскольку требует обработки изображения в высоком разрешении. В то же время, для эффективного обнаружения текста в сложных сценах, снятых на камеру мобильного устройства, нейросетевые модели должны обладать большим рецептивным полем. Как следствие, существующие модели содержат большое количество сверточных слоев и обучаемых параметров.

Для решения проблемы размера моделей поиска текста предлагается компактная архитектура нейронной сети типа Хаф-энкодер [3], основанная на применении быстрого преобразования Хафа (БПХ) в качестве слоя (Рис. 1). Благодаря этому она способна эффективно работать с глобальными признаками в виде прямых линий на картах признаков, что соответствует структуре текстовых строк. Архитектура модели является адаптированной версией архитектуры модели HED-MRZ [4], разработанной для детектирования машиночитаемой зоны (МЧЗ) в документах. Предлагаемая модель содержит всего 31 тысячу обучаемых параметров и весит 116 килобайт, что делает ее самым компактным нейросетевым детектором текста. При этом качество работы сопоставимо с гораздо более тяжелыми моделями, такими как детектор текста фреймворка PaddleOCR [5] на основе DBNet [6], а также CRAFT [7] и специализированный детектор текста для документов MULDT [8].

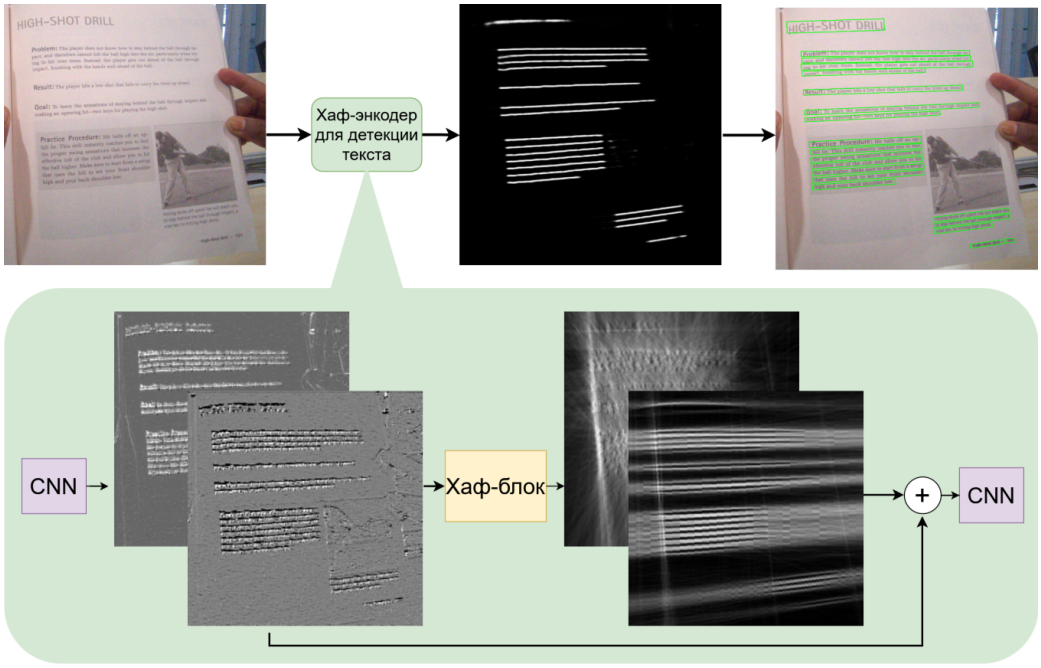


Рис. 1. Предлагаемая модель поиска текста основана на выделении признаков с линейной структурой в пространстве Хафа, соответствующих прямым текстовых строкам на исходном изображении. Для первичного построения локальных признаков и формирования итоговой карты текста используются последовательности сверточных слоев (CNN).

2. ОБЗОР ЛИТЕРАТУРЫ

Методы детекции текста наиболее активно развиваются последние 10 лет, с момента как сверточные нейронные сети продемонстрировали значительное преимущество перед алгоритмами на основе классических методов анализа изображений. Современные исследования сосредоточены на наиболее сложных сценах в естественной среде (англ. “in the wild”), включая изогнутый текст на плакатах и вывесках. Детекторы на основе сегментации, такие как CRAFT [7], FAST [9], DBNet [6] и его улучшенная версия DBNet++ [10], предсказывают тепловые карты текста, на основе которых затем формируются рамки строк. На высокопроизводительных видеокартах они обеспечивают работу в реальном времени, в то время как их производитель-

ность на CPU, особенно мобильных, значительно хуже. Другие методы, основанные на прямом предсказании координат и размеров ограничивающих рамок, такие как TextBPN++ [11] и DPText-DETR [12], используют трансформерный модуль. Такие модели демонстрируют высокое качество поиска текста в сложных сценах, но требуют больших вычислительных ресурсов и много памяти: размер таких моделей может достигать сотен мегабайт.

Несмотря на значительный прогресс в задаче детектирования текста в сложных сценах, количество работ по теме поиска текста на изображениях документов крайне мало. Эта задача во многом считается решенной, поскольку подходы, разрабатываемые для сложных сцен, будут также успешно работать и в случае документов, что отмечено в работе [13]. Однако проблема такого подхода заключается в том, что он не является вычислительно оптимальным: для более простой задачи поиска текста на документах используемые модели могут быть значительно упрощены. Поэтому в работе [14] предлагается быстрый и легковесный детектор текста BusiNet, вдохновленный CRAFT. По сравнению с CRAFT, BusiNet работает в 3 раза быстрее. В схожей работе [8] представлен детектор текста MULDT, ориентированный на изображения документов. Модель имеет небольшой размер в 890 килобайт и демонстрирует результаты, сопоставимые с более тяжелыми аналогами, как для сканированных изображений, так и для фотографий.

Современные архитектуры содержат большое количество сверточных слоев не только для повышения обобщающей способности сети, но и для увеличения рецептивного поля для лучшего захвата контекста. В качестве альтернативы такому экстенсивному подходу можно использовать интегральные операторы в качестве слоев нейронной сети, обладающие глобальным рецетивным полем: например, преобразование Хафа [15, 16]. Этот алгоритм позволяет идентифицировать прямые линии на изображении проецируя его в пространство, где координаты точки соотвествуют конкретной прямой на исходном изображении, а значение в точке соответствует сумме пикселей, лежащих на этой прямой. Поскольку текстовую строку можно рассматривать как набор проходящих через нее прямых линий, алгоритмы на основе преобразования Хафа активно использовались для детектирования текста на изображениях в 1990-2000 годах [17], [18], [19], [20]. В настоящее время быстрое преобразование Хафа используется в качестве слоя в нейронных сетях для значительного упрощения модели и повышения ее качества в тех задачах, где размер рецептивного поля играет первоочередную роль [3], [21], [22], [23], [24]. В работе [4] детектирование машиночитаемой зоны (МЧЗ) было улучшено за счет добавления быстрого преобразования Хафа в качестве слоя в нейронную сеть. Эта задача наиболее близка к задаче поиска текста, поскольку МЧЗ состоит из двух или трех длинных текстовых строк с особой структурой.

3. ПРЕДЛАГАЕМАЯ МОДЕЛЬ

Анализ исследований в задаче поиска текста показывает, что современные модели в первую очередь разработаны для работы в естественной среде (“in the wild”). В более простой предметной области — поиске текста на изображениях документов — эти модели являются избыточными, что затрудняет их применение на устройствах с ограниченными ресурсами. В качестве решения данной проблемы предлагается модель на базе архитектуры Хаф-энкодер [3], которая эффективно выделяет глобальные признаки с линейной структурой с помощью быстрого преобразования Хафа. Такой подход позволяет значительно уменьшить глубину и количество параметров сети, сохранив при этом большое рецептивное поле, необходимое для локализации текстовых строк.

Быстрое преобразования Хафа, используемое в соответствующих слоях нейронной сети, применяется к каждому каналу карт признаков по отдельности. Изначально сам алгоритм был предложен М.Л. Брейди [16] в качестве дискретной аппроксимации преобразования Радона,

которое задается формулой:

$$R(r,\theta) = \int_{-\infty}^{\infty} f(r\cos\theta - z\sin\theta, r\sin\theta + z\cos\theta)dz \tag{1}$$

тогда как быстрое преобразование Хафа выражается следующей формулой:

$$S^{(s,t)} = \sum_{i=0}^{n-1}(I^{(s,t)}(i)) \tag{2}$$

где $I^{(s,t)}$ — интенсивность i-го пикселя, принадлежащего прямой с параметрами (s,t); $S^{(s,t)}$ — Хаф-образ этой прямой. В параметризации (s,t) прямые описываются с помощью s — сдвига линии, и t — меры наклона линии.

Архитектура модели представлена на рисунке 2. Предлагаемая модель рассматривается как расширение модели HED-MRZ [4], разработанной для детектирования машиночитаемой зоны (МЧЗ), на более сложную задачу детектирования текста. Центральным модулем модели является Хаф-блок (рис. 3). Предшествующие ему сверточные слои используются для формирования локальных признаков текстовых фрагментов, которые будут интерпретированы как текстовые строки на основе комбинации Хаф-образов.

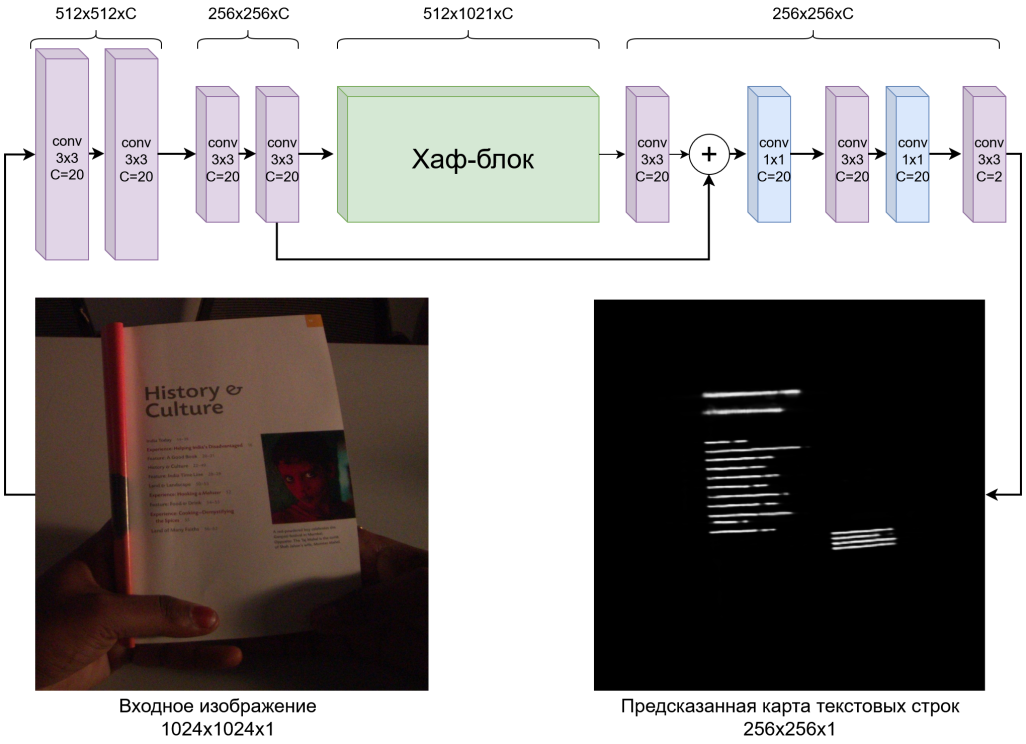


Рис. 2. Архитектура модели Хаф-энкодер для задачи поиска текстовых строк на изображении документа.

Хаф-блок устроен следующим образом. Сперва быстрое преобразование Хафа выполняет отображение карты признаков из пространства изображения (x,y) в параметрическое пространство (s,t) (рис. 4). Яркие пики на Хаф-образе соответствуют наиболее выраженным прямым на изображении. Затем к картам признаков в пространстве Хафа применяются три сверточных слоя с размером ядра 3×3 (рис. 5). Наконец, чтобы преобразовать Хаф-образы

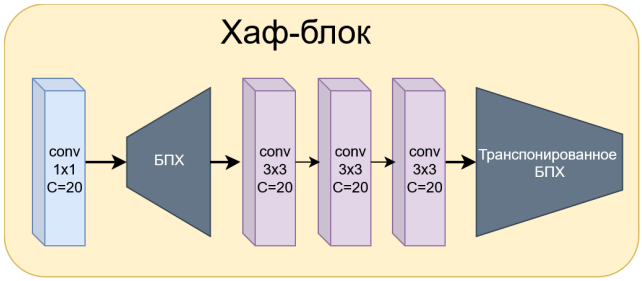


Рис. 3. Хаф-блок работает с глобальными линейными признаками, что позволяет расширить рецептивное поле до размера входного изображения.

Рис. 4. Быстрое преобразование Хафа формирует образ изображения в параметрическом пространстве, где значение в каждой точке соответствует сумме значений пикселей вдоль определенной прямой на входном изображении.

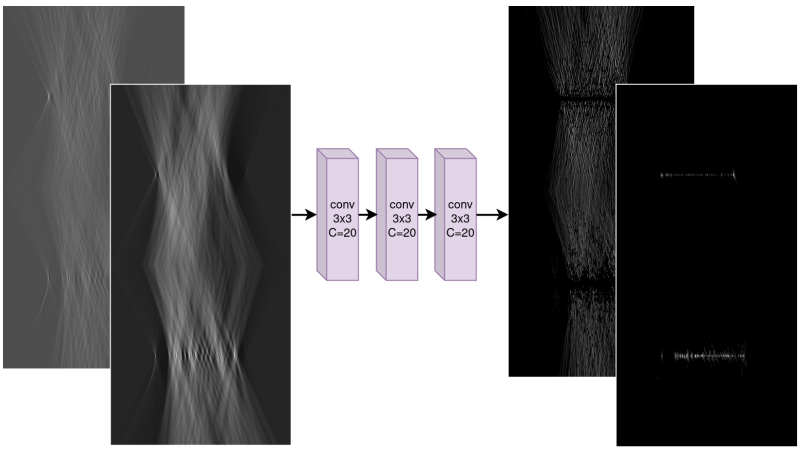


Рис. 5. Сверточные слои, примененные в пространстве Хафа, не только фильтруют карты признаков, но также формируют новые глобальные признаки.

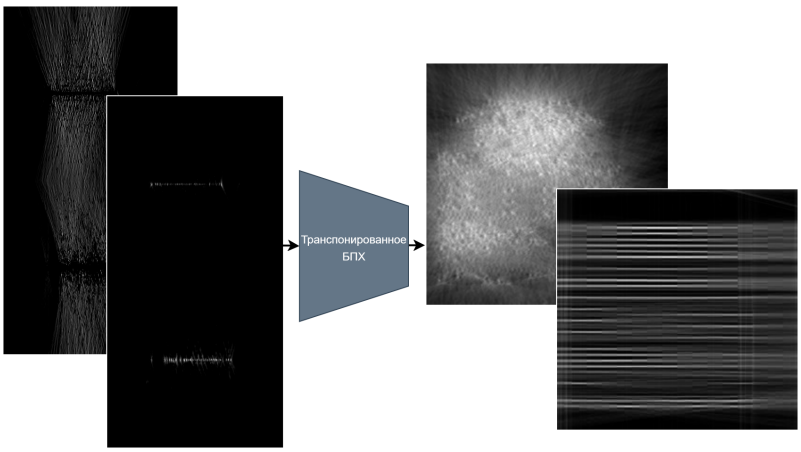


Рис. 6. Транспонированное преобразование Хафа используется для преобразования карт признаков из параметрического пространства (s, t) в пространство изображений (x, y).

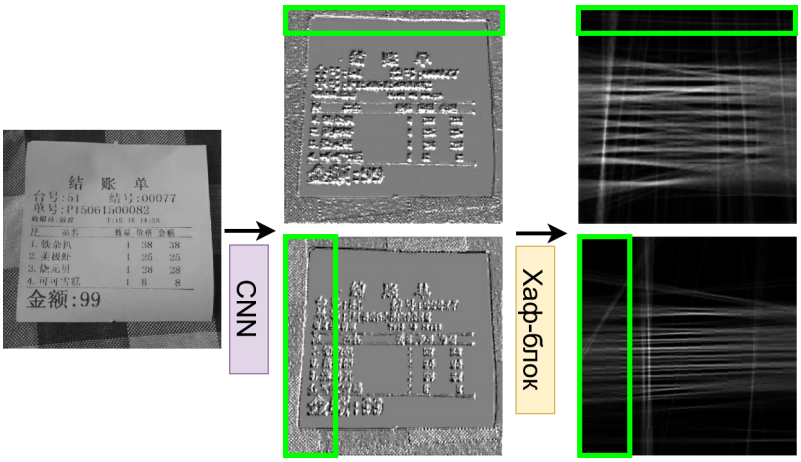


Рис. 7. Хаф-блок фильтрует линии на картах признаков, не связанные с текстовыми строками (выделены зеленой рамкой).

обратно из пространства (s, t) в пространство (x, y), применяется транспонированное быстрое преобразование Хафа [3] (рис. 6).

Хаф-блок фильтрует линии, явно не относящиеся к тексту, как показано на рисунке 7. Затем результат Хаф-блока объединяется с картами признаков с ранних сверточных слоев для комбинирования глобальных и локальных признаков. Следующие четыре сверточных слоя формируют итоговую тепловую карту текста, учитывая как контекст сцены, так и статистическую информацию вдоль прямых. Предлагаемая модель Хаф-энкодера состоит из 31 тысячи обучаемых параметров и имеет размер в 116 килобайт (без квантования). Малый размер позволяет ее легко интегрировать в системы распознавания документов в средах с ограниченными ресурсами.

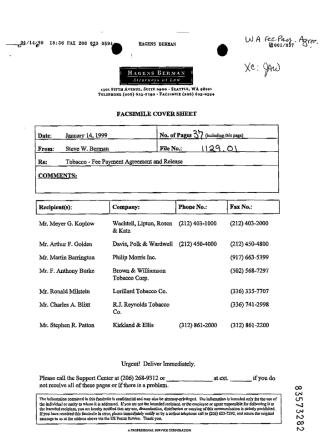
4. ЭКСПЕРИМЕНТ

4.1. Обучение модели

Модель обучалась по методологии, описанной в работе [8]. В качестве обучающих данных использовались изображения с текстом двух типов. Первый тип — простая печать случайных текстовых строк на разных языках на однородных фонах, соответствующих документам: ли-

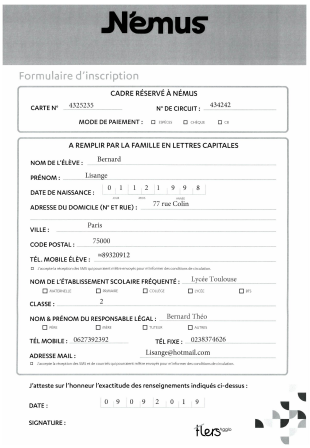
сты бумаги и фоны из удостоверений личности. Второй тип — синтезированные изображения документов, созданные с помощью генератора SDL [25]. Затем изображения делятся на фрагменты размером 1024×1024 пикселей для обучения. Предлагаемая модель обучалась в течение 45 эпох с использованием оптимизатора Adam со скоростью обучения 0.001, коэффициентом L2 регуляризации равным 0.00001, и размером пакета (batch size) равным 48.

4.2. Наборы данных и метрики



(a) FUNSD

(b) SROIE



(c) XFUND

(d) SVRD

Рис. 8. Примеры изображений документов из тестовых наборов данных.

Для оценки качества работы детекторов текста использовался метод TedEval [26]. Он вычисляет точность (Precision, P), полноту (Recall, R) и гармоническое среднее (H-mean, H). Предложенная модель была протестирована на 4 открытых наборах данных:

1. FUNSD [27] — 199 изображений отсканированных форм на английском языке;
2. XFUND [28] — 1043 формы в высоком качестве на 7 языках: китайском, японском, испанском, французском, итальянском, немецком и португальском;
3. SROIE [29] — 986 отсканированных изображений чеков на английском языке;

4. SVRD [30] — 1879 и 1807 изображений в частях task1 и task2 соответственно. Есть как сканированные изображения, так и фотографии. В качестве документов используются чеки, сертификаты, формы, финансовые документы и лицензии. Большинство документов на китайском, также есть документы на английском;

Большинство упомянутых наборов данных разделены на обучающую и тестовую выборки. Так как для обучения модели использовались только синтезированные изображения с текстом, все наборы данных были целиком использованы для тестирования моделей. Примеры изображений из наборов данных показаны на рисунке 8.

4.3. Результаты

Сравнение предложенной модели было проведено со следующими методами поиска текста:

1. CRAFT [7] — модель поиска текста на основе сегментации: на изображении выделяются центры символов и связи между соседними символами, с помощью которых затем формируются рамки слов. Для экспериментов использовалась предобученная модель craft_mlt_25k.pth из официального репозитория на GitHub. CRAFT использует VGG-16 в качестве основной архитектуры для извлечения признаков. Размер модели составляет 79.3 МБ.
2. DBNet [6] — легковесная модель с модулем дифференцируемой бинаризации. Обучена на наборе данных для детектирования текста в естественной среде MSRA-TD500. DBNet использует ResNet-18 для извлечения признаков. Размер модели составляет 52.8 МБ.
3. DBNet++ [10] — улучшенная версия DBNet с адаптивным слиянием масштабов (Adaptive Scale Fusion, ASF). Модель обучена на наборе данных MSRA-TD500. DBNet++ использует ResNet-18 для извлечения признаков. Размер модели составляет 53.3 МБ.
4. DocTr (v. 0.7.0) [31] — фреймворк распознавания изображений документов с открытым исходным кодом. В качестве детектора текста используется модель DBNet с архитектурой ResNet-50, дообученная на изображениях документов. Размер модели составляет 97.2 МБ.
5. PaddleOCR [5] — быстрый и легкий фреймворк распознавания текста с открытым исходным кодом, поддерживающий множество языков. В качестве детектора текста используется DBNet-подобная модель, обученная на приватных данных с применением метода дистилляции знаний. PaddleOCR использовался только в режиме поиска текста; в экспериментах применялись модели PP-OCRv3 (ch, v. 2.7.1) и PP-OCRv4 (ch, v. 2.9.1). Размеры моделей составляют 3.6 МБ и 4.7 МБ соответственно.
6. MULDT [8] — сверхкомпактный мультиязычный детектор текста, предназначенный для документов. Основан на упрощенной сверточной архитектуре с пирамидой масштабов. Размер модели составляет 0.89 МБ.

Модели на основе архитектуры трансформер не рассматриваются, поскольку их размер и вычислительная сложность слишком велики для мобильных устройств и веб-приложений: например, размер модели DPText-DETR [12] составляет 522.8 МБ.

На этапе инференса предложеннной модели входное изображение масштабируется: большая сторона приводится к размеру 1024 пикселя (если она превышает это значение). Затем изображение дополняется нулями до размера 1024×1024 пикселей.

Результаты детекции текста на фотографиях и сканированных изображениях различных документов, таких как чеки, финансовые документы, лицензии и сертификаты, представлены в таблице 1. Результаты для сканированных форм приведены в таблице 2.

5. АНАЛИЗ МОДЕЛИ

Чтобы оценить эффект от выделения глобальных признаков в пространстве Хафа, была обучена аналогичная модель без слоев быстрого преобразования Хафа. Результаты приведены

Таблица 1. Результаты детектирования текста на фотографиях и сканированных изображениях различных документов: чеков, финансовых документов, сертификатов и лицензий.

Модель	Размер (MB)	SVRD task 1			SVRD task 2			SROIE (train)			SROIE (test)		
		P	R	H	P	R	H	P	R	H	P	R	H
DocTr	97.2	0.776	0.445	0.565	0.775	0.438	0.56	0.935	0.912	0.923	0.937	0.915	0.926
CRAFT	79.3	**0.926**	0.854	0.889	0.93	0.862	0.895	**0.957**	0.918	0.937	**0.957**	0.917	0.937
DBNet++	53.3	0.891	0.754	0.817	0.896	0.762	0.823	0.888	0.853	0.871	0.892	0.856	0.874
DBNet	52.8	0.901	0.598	0.719	0.908	0.612	0.731	0.914	0.892	0.903	0.915	0.891	0.903
PaddleOCR v4	4.7	0.89	**0.954**	0.921	0.891	**0.958**	0.923	0.966*	0.983*	0.975*	0.966*	0.979*	0.973*
PaddleOCR v3	3.6	0.914	0.949	**0.931**	0.916	0.951	**0.934**	0.965*	0.973*	0.969*	0.965*	0.979*	0.972*
MULDT	0.89	0.917	0.918	0.917	**0.921**	0.921	0.921	0.933	0.952	**0.943**	0.936	0.952	**0.944**
Хаф-энкодер	**0.116**	0.906	0.93	0.918	0.911	0.934	0.922	0.906	**0.959**	0.931	0.912	**0.959**	0.935

* набор данных SROIE присутствовал в обучающей выборке детектора текста PaddleOCR (ch) [5]

Таблица 2. Результаты детектирования текста на отсканированных формах.

Модель	Размер (MB)	FUNSD (train)			FUNSD (test)			XFUND (train)			XFUND (test)		
			R	H	P	R	H	P	R	H	P	R	H
DocTr	97.2	0.942	0.887	0.914	0.945	0.861	0.901	0.931	0.565	0.703	0.930	0.542	0.684
CRAFT	79.3	**0.974**	0.954	0.964	**0.984**	0.976	**0.979**	**0.975**	0.948	0.961	**0.973**	0.945	**0.959**
DBNet++	53.3	0.913	0.847	0.879	0.928	0.794	0.856	0.945	0.794	0.863	0.944	0.795	0.863
DBNet	52.8	0.960	0.898	0.928	0.969	0.880	0.922	0.969	0.880	0.922	0.967	0.880	0.922
PaddleOcr v4	4.7	0.929	0.960	0.944	0.856	0.957	0.904	0.940	**0.975**	0.957	0.932	**0.975**	0.953
PaddleOcr v3	3.6	0.954	0.962	0.958	0.915	0.960	0.937	0.961	0.964	**0.962**	0.953	0.960	0.956
MULDT	0.89	0.945	0.964	0.954	0.916	0.974	0.944	0.949	0.963	0.956	0.945	0.961	0.953
Хаф-энкодер	**0.116**	0.960	**0.974**	**0.967**	0.925	**0.979**	0.951	0.937	0.939	0.938	0.935	0.941	0.938

в таблице 3. Модель со слоями БПХ превосходит базовую модель на сложных наборах данных, содержащих в том числе фотографии. В то же время, базовая модель показывает схожие или лучшие результаты на наборах данных XFUND и FUNSD, состоящих из сканированных изображений форм. Это можно объяснить тем, что изображения в этих наборах данных имеют высокое качество, размер текста практически фиксирован, а фон документов простой: для решения задачи в таких условиях достаточно локальных признаков. В процессе обучения модель на основе Хаф-энкодера демонстрирует лучшую сходимость по сравнению с базовой моделью. График сходимости представлен на рисунке 9.

Чтобы оценить, насколько предлагаемая модель полагается на признаки извлеченные в пространстве Хафа, выход Хаф-блока был заполнен постоянным значением K, после чего качество модели было измерено на тестовых наборах данных. Особое внимание было уделено сканированным изображениям, где базовая модель без слоев быстрого преобразования Хафа показывает сопоставимое качество. Так как в архитектуре модели есть проброс сигнала (skip connection) в обход Хаф-блока, то возможна ситуация, когда результат Хаф-блока не будет оказывать существенного влияния на результат детекции. В качестве константы K использовались следующие значения:

1. $K = 0$. Моделирование ситуации, когда на карте признаков нет прямых, поэтому выход Хаф-блока заполнен нулями.
2. $K = \max(H)$, где H — выходная карта признаков транспонированного преобразования Хафа. Моделирование ситуации, когда выход Хаф-блока подсвечивает всю карту признаков, указывая на наличие прямых.

Качество модели оказалось около нулевым на всех наборах данных для обеих констант. Это свидетельствует о том, что модель в значительной степени полагается на извлекаемые глобальные признаки.

Примеры детекции текста предлагаемого Хаф-энкодера представлены на рисунке 10. Большинство ошибок выражены в виде ложных срабатываний на сложном фоне, что также видно

Таблица 3. Сравнение качества предложенной модели (Хаф-энкодер) и ее вариации без слоев быстрого преобразования Хафа (Базовая модель).

Набор данных	Модель	P	R	H
SVRD task 1	Хаф-энкодер	**0.906**	**0.93**	**0.918**
	Базовая модель	0.902	0.914	0.908
SVRD task 2	Хаф-энкодер	**0.911**	**0.934**	**0.922**
	Базовая модель	0.906	0.916	0.911
SROIE (train)	Хаф-энкодер	**0.906**	**0.958**	**0.931**
	Базовая модель	0.884	0.932	0.907
SROIE (test)	Хаф-энкодер	**0.912**	**0.959**	**0.935**
	Базовая модель	0.888	0.931	0.909
FUNSD (train)	Хаф-энкодер	0.960	**0.974**	0.967
	Базовая модель	**0.974**	0.963	**0.968**
FUNSD (test)	Хаф-энкодер	0.925	**0.979**	0.951
	Базовая модель	**0.967**	0.976	**0.971**
XFUND (train)	Хаф-энкодер	0.937	0.939	0.938
	Базовая модель	**0.950**	**0.950**	**0.950**
XFUND (val)	Хаф-энкодер	0.935	0.941	0.938
	Базовая модель	**0.947**	**0.942**	**0.944**

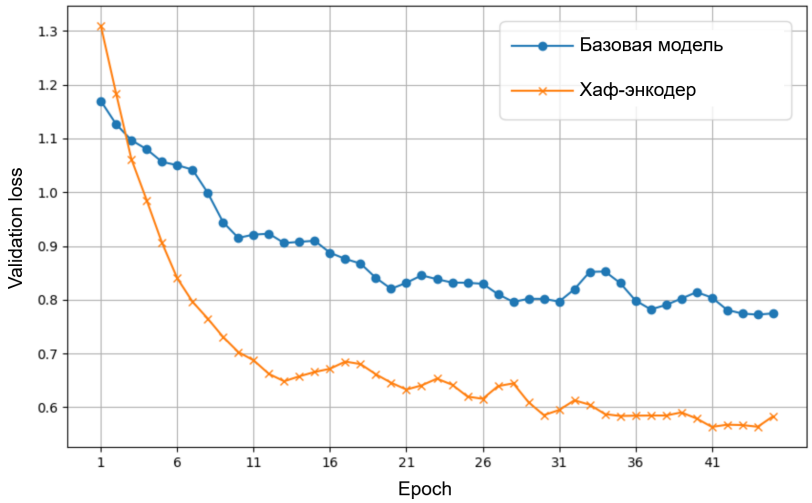


Рис. 9. Сравнение сходимости предложенной модели на основе Хаф-энкодера и ее варианта без слоев Хафа (базовая модель).

из результатов качества детекции текста на публичных наборах данных по метрике точности (Precision, P).

6. ЗАКЛЮЧЕНИЕ

Современные детекторы текста ориентированы на поиск текста в сложных сценах в естественной среде, однако существует более простой, но крайне важный домен — детектирование текста на изображениях документов. Учитывая распространенность систем распознавания и анализа документов, оптимизация моделей существенно улучшает как их практическую применимость, так и возможность запуска на широком спектре устройств. Это особенно важно для мобильных устройств и сред с ограниченными ресурсами. Размер моделей критичен для

Рис. 10. Примеры работы предлагаемой модели на основе архитектуры Хаф-энкодер на реальных изображениях документов из тестовых наборов данных.

их использования в веб-приложениях на базе WASM, где система распознавания документов загружается на устройство при каждом обращении к веб-сервису, поскольку браузер может очищать кэш веб-приложения.

В данной работе предложен сверхкомпактный детектор текста для изображений документов, размер которого позволяет использовать его на любом вычислительном устройстве, даже с экстремальными ограничениями ресурсов. Модель состоит всего из 31 тысячи обучаемых параметров, а ее размер составляет 116 килобайт. Значительное упрощение архитектуры модели стало возможным за счет добавления Хаф-блока, который позволяет увеличить рецептивное поле сети без увеличения количества слоев. При этом качество предложенной модели сопоставимо с современными компактными детекторами текста, такими как DBNet в фреймворке PaddleOCR и детектор текста для документов MULDT.

Основной проблемой представленной модели является количество арифметических операций, несмотря на существенно меньшее число обучаемых параметров. Предлагаемая модель требует 15.4 GFLOPS для обработки изображения размером 1024×1024 пикселей, в то время как MULDT необходимо всего 3.4 GFLOPS. Это обусловлено большей площадью карт признаков в пространстве Хафа, что критично для времени работы сверточных слоев. Для решения этой проблемы необходимо исследовать алгоритмы более компактного представления Хаф-образа [32]. Полный Хаф-образ избыточен, а корректный алгоритм уменьшения размерности Хаф-образа позволит существенно оптимизировать модель. Также проведенные эксперименты показали, что в случае сканированных документов без наклона, с белым фоном и преимущественно единым масштабом достаточно даже очень простой сверточной сети. Однако в случае сложных наборов данных, где присутствуют фотографии с проективными искажениями документа, модель с Хаф-слоями демонстрирует стабильное улучшение.

СПИСОК ЛИТЕРАТУРЫ

1. V.L. Arlazarov and O.A. Slavin. Issues of recognition and verification of text documents. *ITiVS*, 3:55–61, 2023.
2. A.V. Samarin, V.A. Malykh, and P.S. Kalaidin. Id verification method using limited image area. In *Proceedings of the institute for systems analysis russian academy of sciences*, volume 70, pages 15–23, 2020.
3. A. Sheshkus, D.P. Nikolaev, and V.L. Arlazarov. Houghencoder: Neural network architecture for document image semantic segmentation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1946–1950, 2020.
4. D.M. Ershova, A.V. Gayer, A.V. Sheshkus, and V.V. Arlazarov. An ultra-lightweight approach for machine readable zone detection via semantic segmentation and fast hough transform. In Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng, editors, *ICDAR 2024*, volume 14807 of *Lecture Notes in Computer Science (LNCS)*, pages 359–374, Switzerland, 2024. Springer Nature Group.
5. Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoyue Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *ArXiv*, abs/2206.03001, 2022.
6. Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11474–11481, Apr. 2020.
7. Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9357–9366, 06 2019.
8. A.V. Gayer and V.V. Arlazarov. Muldt: Multilingual ultra-lightweight document text detection for embedded devices. *IEEE Access*, 12:170530–170540, 2024.
9. Zhe Chen, Jiahao Wang, Wenhai Wang, Guo Chen, Enze Xie, Ping Luo, and Tong Lu. Fast: Faster arbitrarily-shaped text detector with minimalist kernel representation. In *arXiv*, 2021. 2111.02394.
10. Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. Real-time scene text detection with differentiable binarization and adaptive scale fusion. In *arXiv*, 2022. 2202.10304.
11. Shi-Xue Zhang, Xiaobin Zhu, Chun Yang, and Xu-Cheng Yin. Arbitrary shape text detection via boundary transformer. *IEEE Transactions on Multimedia*, 26:1747–1760, 2022.
12. Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Bo Du, and Dacheng Tao. Dptext-detr: towards better scene text detection with dynamic points in transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI Press, 2023.
13. Krzysztof Olejniczak and Milan Šulc. Text detection forgot about document ocr. In *arXiv*, 2023. 2210.07903.
14. Oshri Naparstek, Ophir Azulai, Daniel Rotman, Yevgeny Burshtein, Peter Staar, and Udi Barzelay. Businet – a light and fast text detection network for business documents. In *arXiv*, 2022. 2207.01220.
15. P. V. C. Hough. Machine Analysis of Bubble Chamber Pictures. *Conf. Proc. C*, 590914:554–558, 1959.
16. Martin L. Brady. A fast discrete approximation algorithm for the radon transform. *SIAM Journal on Computing*, 27(1):107–119, 1998.
17. Sargur N. Srihari and Venu Govindaraju. Analysis of textual images using the hough transform. *Machine Vision and Applications*, 2:141–153, 1989.
18. L. Likforman-Sulem, A. Hanimyan, and C. Faure. A hough based algorithm for extracting text lines in handwritten documents. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 2, pages 774–777 vol.2, 1995.

19. G. Louloudis, B. Gatos, and C. Halatsis. Text line detection in unconstrained handwritten documents using a block-based hough transform approach. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 599–603, 2007.

20. G. Louloudis, B. Gatos, I. Pratikakis, and C. Halatsis. Text line detection in handwritten documents. *Pattern Recognition*, 41(12):3758–3772, 2008.

21. Alexander Sheshkus, Anastasia Ingacheva, Vladimir Arlazarov, and Dmitry Nikolaev. Houghnet: Neural network architecture for vanishing points detection. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 844–849, 2019.

22. Nermin Samet, Samet Hicsonmez, and Emre Akbas. Houghnet: Integrating near and long-range evidence for visual detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4667–4681, 2023.

23. Kai Zhao, Qi Han, Chang-Bin Zhang, Jun Xu, and Ming-Ming Cheng. Deep hough transform for semantic line detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4793–4806, 2022.

24. Lev Teplyakov, Kirill Kaymakov, Evgeny Shvets, and Dmitry Nikolaev. Line detection via a lightweight CNN with a Hough layer. In Wolfgang Osten, Dmitry P. Nikolaev, and Jianhong Zhou, editors, *Thirteenth International Conference on Machine Vision*, volume 11605, page 116051B. International Society for Optics and Photonics, SPIE, 2021.

25. Son Nguyen Truong. Sdl: New data generation tools for full-level annotated document layout. *arXiv*, 2106.15117, 2021.

26. C. Lee, Y. Baek, and H. Lee. Tedeval: A fair evaluation metric for scene text detectors. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 7, pages 14–17, Los Alamitos, CA, USA, sep 2019. IEEE Computer Society.

27. Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6, 2019.

28. Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, and Furu Wei. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3214–3224, Dublin, Ireland, May 2022. Association for Computational Linguistics.

29. Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520, 2019.

30. Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, Yuning Du, Shikun Feng, Xiaoguang Hu, Pengyuan Lyu, Kun Yao, Yuechen Yu, Yuliang Liu, Wanxiang Che, Errui Ding, Cheng-Lin Liu, Jiebo Luo, Shuicheng Yan, Min Zhang, Dimosthenis Karatzas, Xing Sun, Jingdong Wang, and Xiang Bai. Icdar 2023 competition on structured text extraction from visually-rich document images. In Gernot A. Fink, Rajiv Jain, Koichi Kise, and Richard Zanibbi, editors, *Document Analysis and Recognition - ICDAR 2023*, pages 536–552, Cham, 2023. Springer Nature Switzerland.

31. Mindee. doctr: Document text recognition. `https://github.com/mindee/doctr`, 2021.

32. A. Zhabitskaya, A. Sheshkus, and V.L. Arlazarov. Houghtoradon transform: New neural network layer for features improvement in projection space. In Wolfgang Osten, Dmitry Nikolaev, and Johan Debayle, editors, *ICMV 2023*, volume 13072, pages 1307210–1–1307210–8, Bellingham, Washington 98227-0010 USA, Apr. 2024. Society of Photo-Optical Instrumentation Engineers (SPIE).

A compact neural network model for text detection in document images based on the fast Hough transform

A. V. Gayer, A. V. Sheshkus, D. P. Nikolaev, V. V. Arlazarov

Modern document recognition systems are based on neural network models, whose size and computational complexity hinder their deployment on a wide range of devices. This paper proposes an ultra-compact neural network for text detection in resource-constrained environments. It is based on processing global features with a linear structure corresponding to text lines in Hough space. The model size is only 116 kilobytes, which is 8 times smaller than the MULDT text detector and 41 times smaller than the PaddleOCR text detector. Experiments conducted on the FUNSD, SROIE, SVRD, and XFUND datasets show that the proposed model has comparable performance to modern compact text detectors.

KEYWORDS: document recognition, deep learning, fast Hough transform (FHT), text detection.