

---

# Гибридный подход распознавания рукописного текста

---

A Preprint

Войт Руслан Александрович  
МГУ им. М.В. Ломоносова  
Факультет ВМК  
119991, Москва, Ленинские горы, д.1, стр. 52  
s02220048@gse.cs.msu.ru

Местецкий Леонид Моисеевич  
МГУ им. М.В. Ломоносова  
Факультет ВМК,  
119991, Москва, Ленинские горы, д.1, стр. 52  
mestlm@mail.ru

## Abstract

В статье исследуется гибридный подход к распознаванию рукописного текста, сочетающий визуальный анализ растрового изображения и структурный анализ векторного графа штрихов. Существующие растровые модели требуют больших объемов размеченных выборок данных и не учитывают геометрию письма, что снижает их эффективность на исторических документах с высокой вариативностью почерков. Для решения этой проблемы предложена двухпоточная архитектура: визуальный поток на базе Vertical Attention Network дополнен графовым потоком, где граф штрихового разложения строится методом непрерывной скелетизации на основе диаграммы Вороного, а его признаки обрабатываются графовой нейронной сетью с последующим слиянием модальностей и CTC-декодированием. Эксперименты на данных из датасета IAM подтвердили устойчивость предложенного графового представления, продемонстрировав снижение функции потерь на тестовой выборке относительно растровых аналогов. Результаты применимы для оцифровки исторических рукописей и архивных документов.

Keywords Распознавание рукописного текста · Графовые модели · Мультимодальное обучение · Медиальная ось · Диаграмма Вороного

## 1 Введение

Распознавание рукописного текста (Handwritten Text Recognition, HTR) остаётся одной из ключевых задач компьютерного зрения. Особую сложность представляют исторические рукописи с высокой вариативностью почерков, артефактами носителей и дореформенной орфографией, для которых объём размеченных данных зачастую крайне ограничен, а визуальные искажения делают прямое применение стандартных подходов малоэффективным.

Классические подходы рассматривают текст исключительно как растровое изображение. Архитектуры, сочетающие сверточные и рекуррентные слои с функцией потерь CTC (CRNN), а также современные модели на основе архитектуры трансформера (например, TrOCR, HTR-VT, VAN) достигают высоких результатов на крупных эталонных корпусах, таких как IAM. Существенным отличием растровых подходов является их зависимость от больших объемов размеченных данных. Их эффективность существенно снижается в условиях малых выборок: такие модели склонны к переобучению и допускают ошибки на визуально схожих, но структурно различных символах.

В ответ на ограничения растрового подхода в последние годы активно развиваются методы, основанные на анализе траекторий и структуры штрихов: работа InkSight (Lee et al. [2024]) продемонстрировала возможность реконструкции траектории письма из статического изображения с помощью vision-language моделей, а в (Gan et al. [2025]) предложили гибридную архитектуру GCM, объединяющую растровые и графовые признаки через кросс-модальное внимание. Параллельно в (Sharma et al. [2024]) впервые применили графовые нейронные сети к цепочкам кодов, описывающим траектории рукописного ввода, добившись улучшения точности на онлайн- и оффлайн-данных.

Несмотря на успехи, существующие гибридные решения обладают существенными ограничениями. Методы, использующие графовое представление, как правило, опираются на эвристические или нейросетевые процедуры построения графа ( Gan et al. [2025]), которые нестабильны на зашумлённых исторических документах с неравномерной толщиной штрихов и сложной топологией связности. InkSight ( Lee et al. [2024]) не оптимизирован непосредственно для задачи распознавания. Подходы на основе цепочек кодов и GNN ( Sharma et al. [2024]) рассматривают траектории изолированно, без интеграции с растровыми кодировщиками. Таким образом, открытым остаётся вопрос о создании устойчивого, математически обоснованного способа внедрения структурной информации о штрихах в модель распознавания, способную эффективно работать при дефиците размеченных данных.

В настоящей работе предлагается гибридная двухпоточная архитектура, объединяющая модель Vertical Attention Network (VAN), работающую с растровым изображением, и графовую модель на основе штрихового разложения строки. Ключевое отличие предложенного подхода заключается в способе получения графа: он формируется не эвристически, а математически – с помощью аппарата скелетизации на базе диаграммы Вороного многоугольников, разработанного в трудах ( Местецкий [1999, 2009]). Полученный граф обрабатывается графовой нейронной сетью, после чего признаки двух модальностей объединяются через механизмы слияния и декодируются в последовательность символов с помощью CTC-функции потерь.

Эксперименты, проведенные на подмножестве эталонного набора IAM, подтвердили корректность предлагаемого решения: построенные штриховые графы адекватно отражают структуру букв и устойчивы к вариациям начертания. Предложенная модель за счет структурной регуляризации графового потока позволяет достичь более оптимальных значений функции потерь на тестовых данных по сравнению с аналогами, работающими только с растровыми изображениями.

## 2 Обзор литературы

Общедоступные корпуса, такие как IAM ( Marti and Bunke [2002]) и специализированные конкурсы на исторических данных (READ Simistira et al. [2019]) сыграли ключевую роль в стандартизации экспериментов, задав эталонные метрики CER/WER и протоколы разбиения выборок. Современные системы HTR делятся на строчные и страничные; для архивных проектов чаще предпочтительны строчные решения, позволяющие гибко работать со сложной геометрией строк и разнообразием почерков ( Зыков and Местецкий [2025]).

Доминирующим направлением являются глубокие нейронные сети, работающие непосредственно с растровыми изображениями. Такими являются архитектуры, сочетающие свёрточные и рекуррентные слои с CTC-функцией потерь (CRNN) ( Graves et al. [2006], Shi et al. [2015]), а также современные модели на основе трансформеров: TrOCR ( Li et al. [2023]), HTR-VT( Li et al. [2025]), VAN( Зыков and Местецкий [2025]). Такие подходы позволили достичь высоких результатов на стандартных наборах данных. Они рассматривают текст исключительно как визуальный объект и требуют больших объёмов размеченных данных. В условиях работы с историческими коллекциями или малым объёмом разметки визуальные модели склонны к переобучению и часто допускают ошибки, не учитывая структурные различия символов.

В ответ на ограничения растрового подхода в 2024–2025 гг. наметился тренд на возвращение к анализу траекторий и структуры штрихов. В работе InkSight (Lee et al. [2024]) предложили конвертацию офлайн-изображений в цифровые чернила (digital ink) с помощью vision-language модели. InkSight успешно обрабатывает фотографии рукописей с разнородным фоном, однако нацелен на реконструкцию траектории, а не на распознавание текста. В работе FINet ( Zhu et al. [2025]) разработали метод восстановления траекторий китайских иероглифов на основе имитации шрифтов, также без интеграции с распознаванием. Классические геометрические подходы к реконструкции траектории, использующие скелет изображения, были развиты в работе ( Kryzhanovskaya et al. [2020]), где траектория строится как обход скелетного графа. Другие исследования в этом направлении включают применение цепочек кодов (chain codes) для онлайн- и офлайн-рукописи ( Sharma [2013, 2015]) и их комбинацию с рекуррентными сетями ( Singh et al. [2017]).

Параллельно развиваются гибридные модели, объединяющие растровые признаки со структурными. ( Gan et al. [2025]) представили архитектуру GCM (Graph-Enhanced Cross-Modal Mutual Learning), в которой рукописная строка кодируется одновременно как изображение и как граф, построенный путём ресемплинга скелета. Два потока признаков взаимодействуют через кросс-модальное внимание, а взаимная дистилляция декодеров ( $V \rightarrow G$  и  $G \rightarrow V$ ) повышает точность распознавания. На датасетах IAM,

RIMES и ICDAR2013 GCM превосходит чисто визуальные аналоги, что подтверждает эффективность учёта структурной информации. Тем не менее, граф в этой работе строится по эвристическому алгоритму (извлечение скелета, равномерная дискретизация точек), что может быть нестабильно на зашумлённых исторических документах. В (Sharma et al. [2024]) впервые применили графовые нейронные сети к цепочкам кодов, описывающим траектории рукописного ввода, и показали снижение ошибки на MNIST, UNIPEN и GHWT. Однако в их подходе граф формируется напрямую из направленных сегментов цепочек кодов без явного учёта топологии связного письма, а визуальный поток отсутствует. Ряд смежных работ исследуют применение GNN к диаграммам (Wang et al. [2024]) и извлечению информации из рукописных документов (Khanfir et al. [2024]), а также верификацию подписей с помощью графовых трансформеров (jie Yuan et al. [2025]); эти исследования подтверждают потенциал графового представления для анализа рукописного ввода.

Важным источником идей является область распознавания китайских иероглифов, где широко используется декомпозиция на радикалы и штрихи. Методы zero-shot HCCR (Cao et al. [2020], Chen et al. [2021], Zu et al. [2022], Yu et al. [2023a], Luo et al. [2026a]) показывают, что представление символа в виде последовательности радикалов или дерева позволяет обобщать на невидимые классы. Наиболее свежая работа (Luo et al. [2026b]) вводит модуляцию позиционных эмбедингов на основе информационной энтропии радикалов и многоуровневую агрегацию структурных признаков, достигая SOTA в zero-shot сценарии. Близкие по духу исследования используют деревья штрихов (Yu et al. [2023b]), кодирование связей штрихов через трансформеры (Liu et al. [2023]) и динамическую классификацию штрихов (Yang et al. [2023]). Хотя эти работы ориентированы на иероглифику, они убедительно демонстрируют, что явное моделирование иерархической структуры письменных знаков повышает обобщающую способность и устойчивость к вариациям.

Фундаментальный математический аппарат для анализа структуры рукописного текста был разработан в трудах Местецкого и его школы. В (Местецкий [1999, 2009]) предложен метод построения непрерывного скелета бинарного изображения на основе диаграммы Вороного для многоугольной фигуры, обеспечивающий устойчивое геометрическое представление. На этой основе созданы алгоритмы сегментации штрихов, которые разбивают скелетный граф на циклы и цепи, соответствующие каллиграфическим элементам письма, и восстанавливают направление движения пера. Эксперименты на страницах рукописей демонстрируют высокую робастность и скорость работы (2–3 ошибки на строку). Прикладные задачи поиска ключевых слов (Feoktistov and Mestetskiy [2024]) подтверждают практическую ценность штрихового разложения. Тем не менее, до настоящего времени работы по непрерывной морфологии не были интегрированы в гибридные нейросетевые архитектуры для распознавания текста, а существующие гибридные модели опираются на эвристические или нейросетевые процедуры построения графа, которые могут быть нестабильны на сложных исторических данных.

### 3 Математическая постановка задачи

Пусть задано конечное множество изображений рукописных строк  $\mathcal{D} = \{(I^{(n)}, Y^{(n)})\}_{n=1}^N$ , где  $I^{(n)} \in \mathbb{R}^{H \times W \times 3}$  – цветное (или бинаризованное) изображение строки фиксированной высоты  $H$  и ширины  $W$ ,  $Y^{(n)} = (y_1, y_2, \dots, y_{T_n})$  – соответствующая последовательность символов, принадлежащих конечному алфавиту  $\mathcal{A}$  (например, буквы, цифры, знаки пунктуации). Длина последовательности  $T_n$  может варьироваться от примера к примеру.

Требуется построить параметрическую модель  $f_\theta$ , которая по изображению  $I$  восстанавливает последовательность символов:

$$\hat{Y} = f_\theta(I), \quad \hat{Y} \in \mathcal{A}^*,$$

где  $\mathcal{A}^*$  – множество всех конечных последовательностей над алфавитом  $\mathcal{A}$ . Модель  $f_\theta$  представляет собой композицию трёх основных блоков:

- визуальный энкодер  $E_v$ , преобразующий изображение  $I$  в последовательность визуальных признаков  $V = E_v(I)$ ;
- графовый энкодер  $E_g$ , который на основе скелета изображения строит граф штрихов  $G$  и вычисляет структурные признаки  $H = E_g(G)$ ;
- механизм слияния  $\Phi$ , объединяющий признаки  $V$  и  $H$  в общее представление  $U = \Phi(V, H)$ ;
- CTC-декодер  $D$ , который по последовательности  $U$  генерирует условное распределение вероятностей  $P(Y|I; \theta)$  и итоговую последовательность  $\hat{Y}$  путём жадного декодирования или поиска по решётке.

Для обучения модели используется функция потерь CTC (Connectionist Temporal Classification), которая не требует точного выравнивания между признаками и символами. Пусть  $U = (u_1, \dots, u_M)$  – последовательность векторов признаков, полученная после слияния. CTC определяет вероятность целевой последовательности  $Y$  как сумму вероятностей всех возможных выравниваний  $\pi \in \mathcal{A}'^M$  (где  $\mathcal{A}' = \mathcal{A} \cup \{\epsilon\}$ ,  $\epsilon$  – символ «пусто»), которые после приведения удалением повторяющихся символов и  $\epsilon$  дают  $Y$ :

$$P(Y|U) = \sum_{\pi \in \mathcal{B}^{-1}(Y)} \prod_{t=1}^M p_t(\pi_t),$$

где  $p_t(\pi_t)$  – вероятность символа  $\pi_t$  в момент  $t$ , вычисляемая моделью, а  $\mathcal{B}$  – оператор приведения. Логарифмическая функция потерь для одного примера:

$$\mathcal{L}_{CTC}(Y, U) = -\log P(Y|U).$$

Качество распознавания на тестовой выборке оценивается с помощью двух основных метрик:

- символьная частота ошибок (Character Error Rate, CER):

$$CER(Y_{true}, \hat{Y}) = \frac{\rho_{Lev}(Y_{true}, \hat{Y})}{length(Y_{true})} \times 100\%,$$

где  $\rho_{Lev}$  – расстояние Левенштейна (минимальное число вставок, удалений и замен символов);

- словесная частота ошибок (Word Error Rate, WER):

$$WER(Y_{true}, \hat{Y}) = \frac{\rho_{Lev}(W_{true}, \hat{W})}{len(W_{true})} \times 100\%,$$

где  $W_{true}$  и  $\hat{W}$  – последовательности слов, полученные разбиением  $Y_{true}$  и  $\hat{Y}$  по пробелам.

Оптимизационная задача обучения формулируется как минимизация эмпирического риска на обучающем множестве  $\mathcal{D}$ :

$$\theta^* = \arg \min_{\theta} \sum_{n=1}^N \mathcal{L}_{CTC}(Y^{(n)}, U^{(n)}) + \lambda \Omega(\theta),$$

где  $\Omega(\theta)$  – регуляризатор (например,  $L_2$ -норма параметров), а  $\lambda$  – коэффициент регуляризации. Обучение проводится методом стохастического градиентного спуска с использованием алгоритма обратного распространения ошибки.

Решением задачи является обученная модель  $f_{\theta^*}$ , а алгоритмом – процедура прямого прохода по сети для получения  $\hat{Y}$  по входному изображению  $I$ .

## 4 Предложенный метод

В данном разделе подробно описывается гибридная двухпоточная архитектура для распознавания рукописного текста. Общая схема модели представлена на рис. 1. Модель состоит из трёх основных компонентов: визуального энкодера, преобразующего исходное изображение в последовательность визуальных признаков; графового энкодера, извлекающего структурные признаки из штрихового графа; механизма слияния, объединяющего обе модальности с помощью кросс-модального внимания; и CTC-декодера, который по итоговой последовательности признаков генерирует целевую строку текста.

### 4.1 Визуальный поток

В качестве основы визуального потока выбрана архитектура Vertical Attention Network (VAN) (Зыков and Местецкий [2025]), которая показала высокую эффективность при распознавании строк рукописного текста. Входное изображение  $I \in \mathbb{R}^{H \times W \times 3}$  подаётся на вход свёрточной части VAN, состоящей из нескольких этапов понижения разрешения и увеличения числа каналов. На каждом этапе используются блоки с механизмами пространственного вертикального внимания на основе анизотропных сверток, которые позволяют модели фокусироваться на значимых горизонтальных областях, соответствующих

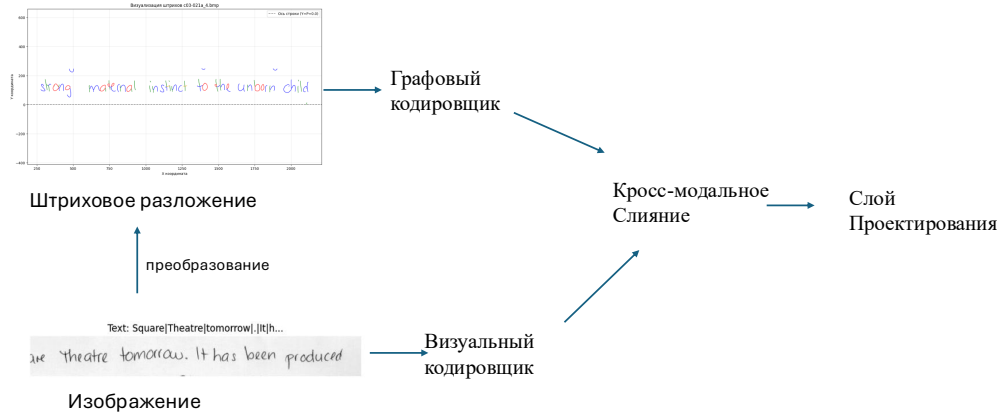


Рис. 1: Общая архитектура предложенной гибридной модели.

расположению символов. Выходом визуального энкодера является последовательность векторных признаков

$$V = (v_1, v_2, \dots, v_M), \quad v_t \in \mathbb{R}^{d_v}, \quad (1)$$

где  $M$  зависит от ширины входного изображения после свёрточных преобразований, а  $d_v$  – размерность признакового пространства (в наших экспериментах  $d_v = 512$ ).

## 4.2 Графовый поток

Графовый поток предназначен для извлечения структурной информации из штрихового представления строки. Его построение выполняется в несколько этапов.

### 4.2.1 Построение графа штрихов на основе непрерывной морфологии

Исходное изображение  $I$  бинаризуется, после чего с помощью алгоритма плоского заметания строится диаграмма Вороного, из которой (Местецкий [1999, 2009]) строится скелет изображения (медиальное представление). Скелет представляет собой геометрический граф, вершинами которого являются точки ветвления и окончания штрихов, а рёбрами – гладкие кривые, соответствующие медиальным осям. Затем выполняется штриховая сегментация (Mestetskiy [2025]), в ходе которой скелетный граф разбивается на элементарные фрагменты – штрихи. Каждый штрих аппроксимирует элементарный фрагмент траектории, соответствующий локально непрерывному движению пера при написании и может быть отнесён к одному из типов (вертикальный отрезок, горизонтальный отрезок, дуга, петля). Отметим, что восстановление траектории пера из оффлайн-изображения является приближённой задачей.

На основе полученного множества штрихов строится граф штрихов  $G = (V, E)$ , где:

- множество вершин  $V = s_1, \dots, s_K$  соответствует отдельным штрихам;
- рёбра  $E \subseteq V \times V$  задаются, если два штриха имеют общую конечную точку в исходном скелете или евклидово расстояние между их ближайшими концевыми точками меньше заданного порога (отражает топологические связи в написании).

Каждой вершине  $s_i$  сопоставляется вектор признаков  $f_i \in \mathbb{R}^{d_f}$ , включающий:

- геометрические характеристики: длина штриха, средняя кривизна, ориентация (угол наклона), координаты центра масс;

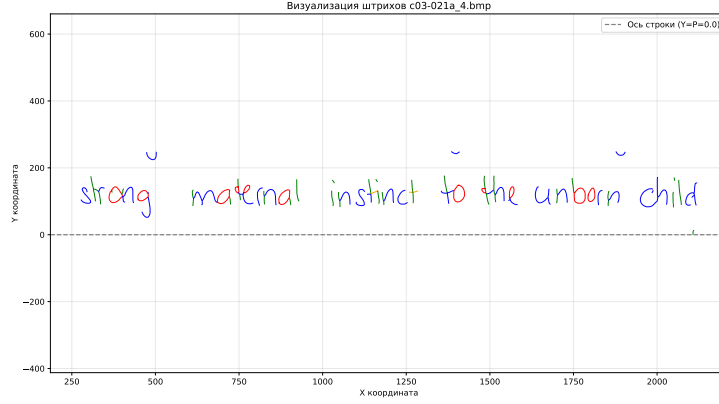


Рис. 2: Пример штрихового разложения рукописного слова: исходное бинаризованное изображение (слева) и построенный на его основе граф штрихов (справа). Различными цветами и типами линий выделены элементарные штрихи.

- тип штриха (кодируется как one-hot вектор);
- параметры, полученные в процессе штрихового разложения (например, радиус вписанной окружности для циклических штрихов).

Таким образом, граф  $G$  полностью описывается матрицей признаков узлов  $X \in \mathbb{R}^{K \times d_f}$  и списком рёбер  $E$ .

#### 4.2.2 Графовый энкодер

Для обработки графа  $G$  используется простая двухслойная графовая нейронная сеть типа GraphSAGE (Hamilton et al. [2018]). Выбор данной архитектуры обусловлен её эффективностью и способностью агрегировать информацию от соседей без необходимости строить полный лапласиан графа. Каждый слой выполняет следующие преобразования:

$$\begin{aligned} h_i^{(1)} &= \text{ReLU}(W_1 \cdot \text{CONCAT}(f_i, \text{MEAN}(f_j : j \in \mathcal{N}(i)))), \\ h_i^{(2)} &= \text{ReLU}(W_2 \cdot \text{CONCAT}(h_i^{(1)}, \text{MEAN}(\{h_j^{(1)} : j \in \mathcal{N}(i)\}))) \end{aligned}$$

где  $\mathcal{N}(i)$  – множество соседей вершины  $i$ ,  $W_1, W_2$  – обучаемые матрицы,  $\text{CONCAT}$  – операция конкатенации,  $\text{MEAN}$  – усреднение по соседям. Выходные эмбединги  $h_i^{(2)} \in \mathbb{R}^{d_h}$  (в наших экспериментах  $d_h = 256$ ) образуют последовательность структурных признаков

$$H = (h_1, h_2, \dots, h_K), \quad h_i \in \mathbb{R}^{d_h}. \quad (2)$$

#### 4.3 Слияние модальностей

Признаки, полученные из визуального и графового потоков, имеют разную природу и размерность. Для их объединения предлагается использовать механизм кросс-модального внимания, который позволяет динамически определять соответствие между пространственными позициями изображения и отдельными штрихами. В отличие от двунаправленных подходов с взаимной дистилляцией (Gan et al. [2025]), мы применяем однонаправленное внимание от визуальной модальности к графовой, что существенно упрощает архитектуру и снижает вычислительную сложность.

Сначала визуальные признаки  $V$  и графовые признаки  $H$  проецируются в единое пространство:

$$\begin{aligned} Q &= VW_Q, \quad W_Q \in \mathbb{R}^{d_v \times d_k}, \\ K &= HW_K, \quad W_K \in \mathbb{R}^{d_h \times d_k}, \\ V_{val} &= HW_V, \quad W_V \in \mathbb{R}^{d_h \times d_v}, \end{aligned}$$

где  $W_Q, W_K, W_V$  – обучаемые матрицы проекций,  $d_k$  – размерность пространства ключей (в наших экспериментах  $d_k = 128$ ). Затем вычисляется матрица внимания

$$A = \text{softmax} \left( \frac{QK^\top}{\sqrt{d_k}} \right) \in \mathbb{R}^{M \times K}, \quad (3)$$

где  $A_{t,i}$  интерпретируется как степень соответствия  $t$ -й пространственной позиции изображения  $i$ -му штриху. Взвешенная сумма графовых признаков для каждой визуальной позиции:

$$\tilde{H} = AV_{val} \in \mathbb{R}^{M \times d_v}. \quad (4)$$

Итоговая последовательность признаков формируется путём конкатенации исходных визуальных признаков и взвешенных графовых:

$$U = [V | \tilde{H}] \in \mathbb{R}^{M \times (d_v + d_v)}, \quad (5)$$

где  $|$  обозначает конкатенацию по размерности признаков. Альтернативно может использоваться аддитивное слияние  $U = V + \tilde{H}$ , что сокращает размерность, но в данной работе мы используем конкатенацию как более общий способ.

На выходе мы применим обучаемый линейный слой для проекции признаков перед декодированием.

$$U' = UW_O, \quad W_O \in \mathbb{R}^{2d_v \times d_v} \quad (6)$$

Отметим, что предложенный механизм слияния является упрощенным и не требует сложной синхронизации модальностей, что делает его удобным для обучения на ограниченных выборках данных.

#### 4.4 Декодирование

Для преобразования последовательности признаков  $U'$  в итоговую последовательность символов  $Y = (y_1, \dots, y_T)$  применяется СТС-декодер. Декодирование реализуется с помощью линейного слоя, преобразующего каждый вектор  $u'_t$  в вектор логитов размера  $|\mathcal{A}'|$ , где  $\mathcal{A}' = \mathcal{A} \cup \{\epsilon\}$  – расширенный алфавит с символом пустоты  $\epsilon$ . Вероятность символа  $k$  в позиции  $t$  вычисляется с помощью softmax:

$$p_t(k) = \frac{\exp((W_{ctc}u'_t)_k)}{\sum_{k' \in \mathcal{A}'} \exp((W_{ctc}u'_t)_{k'})}. \quad (7)$$

Условная вероятность целевой последовательности  $Y$  при заданной последовательности признаков  $U'$  определяется через суммирование по всем возможным выравниваниям  $\pi \in \mathcal{A}'^M$ , которые после приведения (удаления повторяющихся символов и символов  $\epsilon$ ) дают  $Y$ :

$$P(Y | U') = \sum_{\pi \in \mathcal{B}^{-1}(Y)} \prod_{t=1}^M p_t(\pi_t), \quad (8)$$

где  $\mathcal{B}$  – оператор приведения. В качестве функции потерь используется отрицательный логарифм этой вероятности:

$$\mathcal{L}_{CTC}(Y, U') = -\log P(Y | U'). \quad (9)$$

На этапе инференса применяется жадное декодирование, выбирающее на каждом шаге символ с максимальной вероятностью, после чего выполняется приведение и удаление символов  $\epsilon$ .

#### 4.5 Обучение

Модель обучается сквозным образом минимизацией эмпирического риска на обучающем множестве  $\mathcal{D}$ :

$$\mathcal{L}_{total} = \sum_{n=1}^N \mathcal{L}_{CTC}(Y^{(n)}, U^{(n)}) + \lambda \Omega(\theta), \quad (10)$$

где  $\Omega(\theta)$  –  $L_2$ -регуляризация параметров,  $\lambda$  – коэффициент регуляризации. Визуальный энкодер инициализируется предобученными весами VAN, а графовый энкодер и слои слияния – случайно. Оптимизация выполняется методом Adam с начальной скоростью обучения  $10^{-3}$  и уменьшением при достижении плато. Для предотвращения переобучения применяется нормализация и dropout с вероятностью 0.3 после каждого слоя GNN и после слоя внимания.

## 5 Эксперименты

Для оценки эффективности предложенного гибридного подхода была проведена серия экспериментов, направленных на сравнение качества распознавания полной двухпоточной модели с её визуальным аналогом.

### 5.1 Настройка эксперимента

В качестве основы для экспериментов использовалось подмножество данных из публичного датасета IAM. Набор данных был разделен на обучающую (90%) и тестовую (10%) выборки. Все эксперименты проводились с аппаратным ускорением на графическом процессоре NVIDIA Tesla P100. Для реализации моделей использовались библиотеки PyTorch (версия 2.x) и PyTorch Geometric. Из-за вычислительных ограничений и в целях ускорения прототипирования, в качестве визуального энкодера вместо сложной архитектуры VAN использовалась упрощенная сверточная нейронная сеть (CNNEncoder), описанная в листинге кода. Это позволило сфокусироваться на проверке работоспособности основной идеи — слияния модальностей.

Были обучены три конфигурации моделей:

- Гибридная модель (OCRModel): Предложенная двухпоточная архитектура с визуальным и графовым энкодерами и механизмом кросс-модального слияния.
- Базовая модель (OCRModelCNNOnly): Модель, использующая только визуальный поток (CNNEncoder) с последующим линейным классификатором. Служит базовой линией для сравнения.
- Модель на графах (OCRModelGNNOnly): Модель, использующая только графовый поток (GNNEncoder). Её обучение не привело к сходимости из-за сложности выравнивания последовательности узлов графа с текстовой транскрипцией, что подтверждает необходимость визуальной модальности для данной задачи.

Все модели обучались с использованием CTC-потерь и оптимизатора Adam с начальной скоростью обучения  $10^{-3}$  в течение 90 эпох. Размер мини-батча составлял 8. Оценка качества производилась по значению функции потерь на обучающей и тестовой выборках.

### 5.2 Результаты и анализ

Динамика изменения функции потерь в процессе обучения для гибридной и базовой моделей представлена на рис. 3. Для наглядности отображения динамики сходимости на рисунке представлен интервал с 30-й по 90-ю эпоху; исключение начального участка обусловлено необходимостью детального рассмотрения поведения моделей в окрестности оптимума, где различия становятся наиболее показательными. Базовая модель сошла быстрее гибридной, достигнув плато за начальные 30 эпох. Гибридная модель, как видно из графика, продолжает обучение и демонстрирует улучшение на протяжении всего периода обучения. В результате на интервале 60-90 эпох предложенная двухпоточная архитектура показывает более низкие значения функции потерь на тестовой выборке по сравнению с базовой моделью, что свидетельствует о её лучшей обобщающей способности.

Для количественной оценки результатов был проведен анализ средних значений функции потерь на последних 30 эпохах обучения, когда процесс оптимизации вышел на плато. Полученные усредненные значения представлены в таблице 1.

Модель	Train Loss	Test Loss
Базовая модель (CNN)	3.096	3.025
Гибридная модель (CNN + GNN)	3.175	3.009

Таблица 1: Средние значения функции потерь на обучающей и тестовой выборках за последние 30 эпох обучения.

Данные таблицы 1 демонстрируют, что гибридная модель, несмотря на несколько более высокое среднее значение потерь на обучающей выборке (3.175 против 3.096), показывает лучшее обобщение на тестовой выборке. Средний лосс для гибридной модели составил 3.009, что ниже значения 3.025, полученного для базовой модели CNN. Это свидетельствует о меньшей склонности гибридной модели к переобучению.

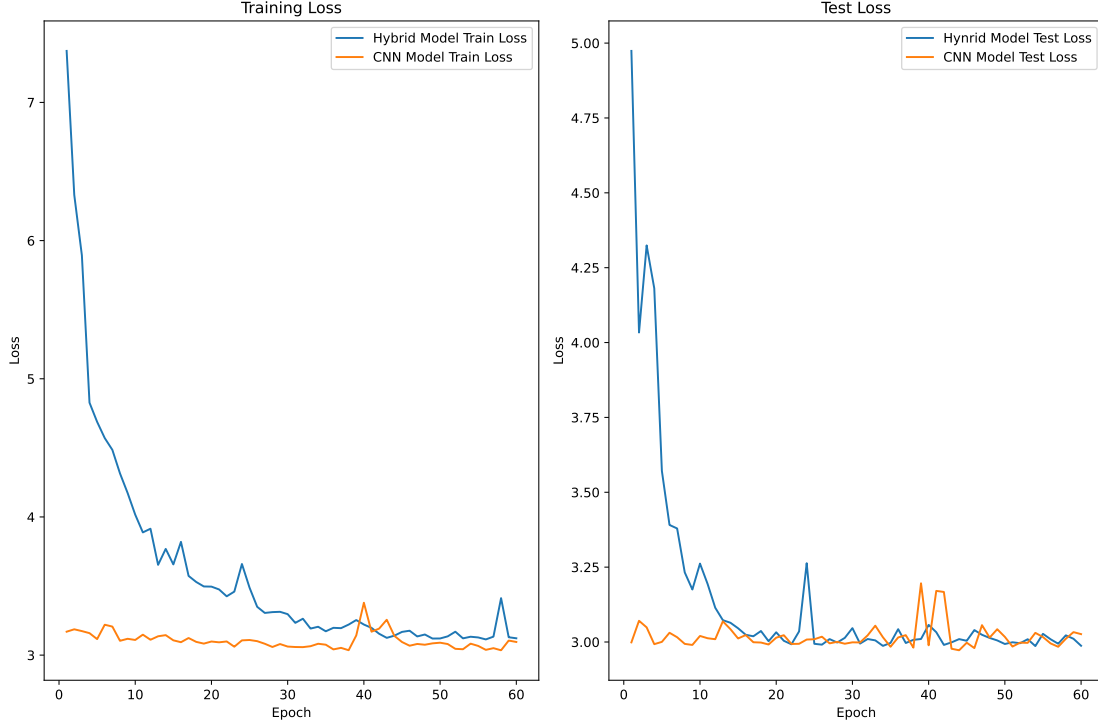


Рис. 3: Динамика функции потерь на обучающей и тестовой выборках для гибридной модели и базовой модели CNN.

Данный эффект можно объяснить регуляризующим воздействием графового потока. Структурная информация о штрихах, извлеченная из графа, выступает в роли дополнительного ограничивающего фактора, который не позволяет визуальному энкодеру подстраиваться под незначительные и нерелевантные шумы в обучающих данных, тем самым улучшая способность модели к обобщению.

Несмотря на упрощенную визуальную архитектуру, использованную в эксперименте, полученные результаты подтверждают основной тезис работы: совместное использование визуальной и структурной информации о рукописном тексте повышает устойчивость модели и качество распознавания по сравнению с чисто визуальными подходами.

## 6 Заключение

В настоящей работе был предложен и исследован гибридный подход к распознаванию рукописного текста, основанный на совместном использовании визуальной и структурной информации. Ключевой особенностью метода является применение математически обоснованного аппарата непрерывной морфологии для построения графа штрихов, который в отличие от эвристических методов обеспечивает устойчивость к вариациям почерка и шумам изображения.

Предложенная двухпоточная архитектура объединяет визуальный энкодер (в качестве которого может выступать современная модель Vertical Attention Network) и графовый энкодер на базе GraphSAGE, обрабатывающий штриховое представление. Слияние модальностей реализовано через механизм однонаправленного кросс-модального внимания, что позволяет динамически устанавливать соответствие между пространственными позициями изображения и элементарными штрихами.

Экспериментальное исследование, проведенное на подмножестве датасета IAM с использованием упрощенной сверточной визуальной модели, подтвердило эффективность предложенного подхода. Гибридная модель продемонстрировала лучшее обобщение на тестовой выборке по сравнению с чисто визуальным аналогом, что свидетельствует о регуляризующем эффекте графового потока и его способности снижать переобучение. Полученные результаты подтверждают гипотезу о том, что структурная информация о

штрихах повышает устойчивость модели распознавания, особенно в условиях ограниченного объема данных.

В качестве направлений дальнейших исследований можно выделить интеграцию в предложенную архитектуру полноценного визуального энкодера VAN, расширение экспериментов на исторические коллекции с вычислением метрик CER/WER, а также исследование возможности использования более сложных механизмов взаимодействия модальностей, включая двунаправленное внимание.

## Список литературы

- C.-Y. Lee, M. Wilber, M. Sun, T. Bui, Y. Yang, Y. Wang, et al. Inksight: Offline-to-online handwriting conversion by learning to read and write. arXiv preprint, 2024. arXiv:2402.05804.
- Ji Gan, Yupeng Zhou, Yanming Zhang, Jiaxu Leng, and Xinbo Gao. Structure-aware handwritten text recognition via graph-enhanced cross-modal mutual learning. In James Kwok, editor, Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, pages 5154–5162. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi:10.24963/ijcai.2025/574. URL <https://doi.org/10.24963/ijcai.2025/574>. Main Track.
- Anuj Sharma, Sukhdeep Singh, and S Ratna. Graph neural network based handwritten trajectories recognition, 2024. URL <https://arxiv.org/abs/2405.09247>.
- Л. М. Местецкий. Скелетизация многоугольной фигуры на основе обобщенной триангуляции Делоне. Программирование, (3):16–31, 1999.
- Л. М. Местецкий. Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры. Физматлит, М., 2009.
- U.-V. Marti and H. Bunke. The iam-database: an english sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition, 5(1):39–46, Nov 2002. ISSN 1433-2833. doi:10.1007/s100320200071. URL <https://doi.org/10.1007/s100320200071>.
- Fotini Simistira, Rajkumar Saini, Derek Dobson, Jon Morrey, and Marcus Liwicki. Icdar 2019 historical document reading challenge on large structured chinese family records. 03 2019. doi:10.48550/arXiv.1903.03341.
- В. П. Зыков and Л. М. Местецкий. Пост-коррекция слабой расшифровки большими языковыми моделями в итерационном процессе распознавания рукописей. Электронные библиотеки, 28(6):1386–1414, 2025.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. volume 2006, pages 369–376, 01 2006. doi:10.1145/1143844.1143891.
- Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, PP, 07 2015. doi:10.1109/TPAMI.2016.2646371.
- Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI Conference on Artificial Intelligence, 37(11):13094–13102, Jun. 2023. doi:10.1609/aaai.v37i11.26538. URL <https://ojs.aaai.org/index.php/AAAI/article/view/26538>.
- Yuting Li, Dexiong Chen, Tinglong Tang, and Xi Shen. Htr-vt: Handwritten text recognition with vision transformer. Pattern Recognition, 158:110967, February 2025. ISSN 0031-3203. doi:10.1016/j.patcog.2024.110967. URL <http://dx.doi.org/10.1016/j.patcog.2024.110967>.
- Yuanping Zhu, Shengnan Li, Hui Wang, and Feilong Wei. Finet: Handwriting trajectory reconstruction of chinese characters based on the font imitate network. Pattern Recognition, 157:110949, 01 2025. doi:10.1016/j.patcog.2024.110949.
- Svetlana Kryzhanovskaya, Sergey Arseev, and Leonid Mestetskiy. Pen trace reconstruction with skeleton representation of a handwritten text image. Proceedings of the 30th International Conference on Computer Graphics and Machine Vision (GraphiCon 2020). Part 2, pages paper27–1, 12 2020. doi:10.51130/graphicon-2020-2-3-27.
- Anuj Sharma. Recovery of drawing order in handwritten digit images. 12 2013. doi:10.1109/ICIIP.2013.6707630.
- Anuj Sharma. A combined static and dynamic feature extraction technique to recognize handwritten digits. Vietnam Journal of Mathematics, 2:133–142, 01 2015. doi:10.1007/s40595-014-0038-1.
- Sukhdeep Singh, Anuj Sharma, and Indu Chhabra. A dominant points-based feature extraction approach to recognize online handwritten strokes. International Journal on Document Analysis and Recognition (IJDAR), 20, 03 2017. doi:10.1007/s10032-016-0279-x.
- Hao-Zhe Wang, Yan-Ming Zhang, Fei Yin, and Lin-Lin Huang. Diaggcn: An end-to-end framework for online handwritten diagram recognition. In 2024 IEEE 8th International Conference on Vision, Image and Signal Processing (ICVISIP), pages 1–5, 2024. doi:10.1109/ICVISIP64524.2024.10959153.

- Yessine Khanfir, Marwa Dhiaf, Emna Ghodhbani, Ahmed Cheikh Rouhou, and Yousri Kessentini. Graph neural networks for end-to-end information extraction from handwritten documents. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 504–512, January 2024.
- Hai jie Yuan, Heng Zhang, and Fei Yin. Online handwritten signature verification based on temporal-spatial graph attention transformer, 2025. URL <https://arxiv.org/abs/2510.19321>.
- Zhong Cao, Jiang Lu, Sen Cui, and Changshui Zhang. Zero-shot handwritten chinese character recognition with hierarchical decomposition embedding. Pattern Recognition, 107:107488, 06 2020. doi:10.1016/j.patcog.2020.107488.
- Jingye Chen, Bin Li, and Xiangyang Xue. Zero-shot chinese character recognition with stroke-level decomposition, 2021. URL <https://arxiv.org/abs/2106.11613>.
- Xinyan Zu, Haiyang Yu, Bin Li, and Xiangyang Xue. Chinese character recognition with augmented character profile matching. In Proceedings of the 30th ACM International Conference on Multimedia, MM '22, page 6094–6102, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi:10.1145/3503161.3547827. URL <https://doi.org/10.1145/3503161.3547827>.
- Haiyang Yu, Xiacong Wang, Bin Li, and Xiangyang Xue. Chinese text recognition with a pre-trained clip-like model through image-ids aligning, 2023a. URL <https://arxiv.org/abs/2309.01083>.
- Qiuming Luo, Tao Zeng, Xuan Wei, and Chang Kong. Radical sequence encoding with fine-tuned clip for handwritten chinese character recognition. In Xu-Cheng Yin, Dimosthenis Karatzas, and Daniel Lopresti, editors, Document Analysis and Recognition – ICDAR 2025, pages 408–424, Cham, 2026a. Springer Nature Switzerland. ISBN 978-3-032-04624-6.
- Qiuming Luo, Tao Zeng, Feng Li, Heming Liu, Rui Mao, and Chang Kong. Entropy-aware structural alignment for zero-shot handwritten chinese character recognition, 2026b. URL <https://arxiv.org/abs/2602.03913>.
- Haiyang Yu, Jingye Chen, Bin Li, and Xiangyang Xue. Chinese character recognition with radical-structured stroke trees. Mach. Learn., 113(6):3807–3827, December 2023b. ISSN 0885-6125. doi:10.1007/s10994-023-06450-6. URL <https://doi.org/10.1007/s10994-023-06450-6>.
- Jing-Yu Liu, Yan-Ming Zhang, Fei Yin, and Cheng-Lin Liu. Transformer-based stroke relation encoding for online handwriting and sketches. Pattern Recognition, 148:110131, 11 2023. doi:10.1016/j.patcog.2023.110131.
- Yu-Ting Yang, Yan-Ming Zhang, Xiao-Long Yun, Fei Yin, and Cheng-Lin Liu. Dygat: Dynamic stroke classification of online handwritten documents and sketches. Pattern Recognition, 141:109564, 04 2023. doi:10.1016/j.patcog.2023.109564.
- Dmitrii Feoktistov and Leonid Mestetskiy. Keywords spotting in russian handwritten documents based on strokes segmentation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-2/W5-2024:49–54, 12 2024. doi:10.5194/isprs-archives-XLVIII-2-W5-2024-49-2024.
- L. Mestetskiy. Stroke segmentation of handwritten text based on medial representation. Pattern Recognition and Image Analysis, 34:1185–1191, 04 2025. doi:10.1134/S1054661824701256.
- William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs, 2018. URL <https://arxiv.org/abs/1706.02216>.