УДК 004.021+004.853

Разработка комплексной методики для аннотирования изображений биопсии

Хрюкин Е.А.[1], Мартыненко Т.В.[1], Васяева Т.А.[1], Швороб Д.С.[2]
[1]ФГБОУ ВО «Донецкий Национальный Технический Университет»
[2]Донецкий государственный медицинский университет им. М. Горького
E-mail: wadealer@yandex.ru

Аннотация:
*В статье предложена комплексная методика аннотирования медицинских изображений биопсии на основе языка **Python** и библиотеки **large-image**, включающая конвертацию изображений в формат **JPEG**, автоматическое выделение участков биопсии, их разбиение на фрагменты и использование инструмента **Makesense.AI** для аннотирования, позволяющая сэкономить вычислительные ресурсы и упростить работу специалистов, обеспечивающая эффективную подготовку данных для обучения нейросетей.*

Введение

Обработка и аннотирование изображений высокого разрешения представляет собой актуальную задачу в сфере анализа медицинских данных, особенно в контексте компьютерной диагностики и подготовки обучающих выборок для нейросетевых моделей.

Такие изображения, как правило, характеризуются высокой детализацией и большим объёмом, что требует использования специализированных алгоритмов и инструментов для эффективной работы с ними [1].

Особую сложность представляет аннотирование гистологических срезов, полученных в результате биопсии, из-за необходимости точного выделения морфологических структур на масштабных данных. Традиционные методы аннотирования оказываются трудоёмкими и малоэффективными при работе с подобными изображениями, особенно в условиях ограниченных вычислительных ресурсов [2].

В настоящей работе сделан акцент на оптимизации этапа подготовки данных за счёт автоматизации предварительной обработки изображений, алгоритмической фильтрации нерелевантных участков и адаптации существующих **open source**-инструментов для нужд конкретной задачи. Анализ современных решений, таких как **SlideRunner**, **QuPath** и специализированных **Python**-библиотек, позволил сформировать обоснованные требования к проектируемому подходу.

Целью статьи является формализация и реализация методики, позволяющей повысить эффективность аннотирования медицинских изображений большого формата при сохранении качества и полноты данных, необходимых для последующего обучения моделей компьютерного зрения.

Под эффективностью аннотирования понимается предоставление патологоанатому инструментария, позволяющего проаннотировать изображение максимально точно, без ошибок. Точность аннотирования в конечном счете может быть определена только экспертно.

Особенности медицинских изображений большого размера

Медицинские изображения большого размера, такие как снимки биопсии, представляют собой сложные данные, которые требуют особого подхода к обработке и анализу. Один снимок биопсии может занимать гигабайты памяти, особенно если он состоит из множества слоёв или сделан с использованием трёхмерной визуализации. Это создаёт сложности при передаче, хранении и обработке данных, а также увеличивает время, необходимое для аннотирования.

Изображения биопсии обычно создаются с использованием мощных микроскопов, что обеспечивает детализированное отображение тканей. Это необходимо для точного анализа микроскопических структур, таких как клетки, сосуды и патологические изменения. Однако высокое разрешение приводит к огромному объёму данных, что усложняет их хранение, обработку и аннотирование [3].

Формат **TIFF** (**Tagged Image File Format**) является одним из наиболее популярных форматов для хранения медицинских изображений благодаря своей гибкости и способности сохранять данные высокого разрешения. Основные преимущества формата **TIFF**: поддержка высокого разрешения, многостраничные файлы, поддержка сжатия. Для работы с **TIFF**-изображениями в программных приложениях используются различные библиотеки, каждая из которых имеет свои особенности и ограничения (таблица 1).

Таблица 1 – Библиотеки для работы с форматом **TIFF**

Наименование библиотеки	**Описание**
Libtiff	Библиотека с открытым исходным кодом, предоставляющая базовые функции для чтения, записи и манипуляции **TIFF**-файлами.
OpenSlide	Специализированная библиотека для работы с цифровыми слайдами, включая медицинские изображения в формате **TIFF.**
Pillow (Python Imaging Library)	Библиотека для работы с изображениями в **Python**, поддерживающая чтение и запись **TIFF**-файлов.
Bio-Formats	Библиотека, предназначенная для работы с биомедицинскими изображениями, включая **TIFF**.
Tifffile	**Python**-библиотека, специально разработанная для работы с **TIFF**-файлами, включая многослойные изображения и сжатие.
GDAL	Библиотека, изначально разработанная для геопространственных данных, поддерживает работу с **TIFF**.

Математическая постановка задачи

Рассматривается задача аннотирования больших медицинских изображений высокого разрешения, где требуется эффективно выделять и объединять области интереса.

Пусть I – изображение, заданное как отображение:

$$I: \Omega \subset \mathbb{R}^2 \to \mathbb{R}^c, \quad (1)$$

где Ω – ограниченная область в двумерном евклидовом пространстве, $\mathbb{R}$ – множество вещественных чисел, а c – количество цветовых каналов.

Требуется найти множество ограничивающих прямоугольников Ri, минимизирующее суммарную ошибку разметки и штраф за перекрытие областей. Формально, задача формулируется как задача оптимизации:

$$\min_{\mathrm{R_i}} \left(\sum_i L_{\text{разметки}}(R_i) + \lambda L_{\text{перекрытия}}(R_i) \right), \quad (2)$$

где $\mathrm{L}_{\text{разметки}}(\mathrm{R_i})$ – функция ошибки разметки для прямоугольника Ri;

$\mathrm{L}_{\text{перекрытия}}(\mathrm{R_i})$ – функция штрафа за перекрытие прямоугольников;

$\lambda > 0$ – регуляризационный коэффициент, который регулирует баланс между точностью разметки и минимизацией перекрытия.

Таким образом, задача состоит в построении эффективной разметки изображения, учитывающей точность выделения областей интереса. Функция ошибки разметки вычисляется по формуле:

$$L_{\text{разметки}}(R_i) = 1 - IoU(R_i, G_i), \quad (3)$$

где G_i – соответствующий эталонный (истинный) прямоугольник;

IoU (Intersection over Union) – пересечение по объединению.

Разработка комплексной методики для аннотирования изображений биопсии

Современные инструменты для аннотирования медицинских изображений предоставляют широкий функционал для работы с данными, однако они имеют ряд ограничений, особенно при работе с изображениями большого размера. Наиболее популярные решения и их недостатки представлены в таблице 2.

Для преодоления описанных выше ограничений, упрощения процесса аннотирования изображений биопсии и решения поставленной задачи была разработана комплексная методика, основанная на использовании языка **Python** и специализированных библиотек.

Методика включает применение следующих инструментов и технологий.

Библиотека **large-image** [4] **–** позволяет эффективно загружать и обрабатывать изображения по частям, что значительно снижает требования к оперативной памяти. Предоставляет унифицированный интерфейс для работы с низкоуровневыми библиотеками, описанными в предыдущем разделе. В качестве бэкэнда выбрана библиотека **tifffile,** которая в ходе экспериментов показала наилучшую производительность.

1. Конвертация изображения в формат **JPEG** с уменьшением разрешения. При этом сохраняется необходимое соотношение между размером файла и качеством изображения, чтобы обеспечить точность аннотации.
2. Для постобработки сконвертированного изображения используется библиотека **OpenCV** [5] – отраслевой стандарт для работы с изображениями.

Обобщенный алгоритм обработки изображения представлен на рис. 1 и включает следующие этапы: бинаризация и поиск контуров; разделение изображения на фрагменты; сохранение фрагментов.

Таблица 2 – Инструменты для аннотирования медицинских изображений.

Название	**Описание**	**Недостатки**
QuPath [6]	Один из наиболее распространённых инструментов для аннотирования биомедицинских изображений. Предоставляет мощные функции для работы с изображениями высокого разрешения.	Ограниченная производительность при обработке сверхбольших изображений, высокая нагрузка на аппаратное обеспечение, сложности в интеграции с пользовательскими скриптами или внешними системами.
ImageJ/Fiji [7, 8]	Широко используемый инструмент для анализа медицинских изображений, предоставляет модульность и поддержку множества плагинов.	Интерфейс, требующий значительного обучения, отсутствие специализированных инструментов для аннотирования биопсийных изображений, сложность обработки больших объёмов данных.
Aperio ImageScope	Ориентирован на просмотр и аннотирование цифровых слайдов.	Ограниченная поддержка форматов изображений, высокая стоимость лицензии, отсутствие гибкости для кастомизации рабочих процессов.
DeepLabCut и аналогичные инструменты	Предназначены для использования в исследованиях, связанных с машинным обучением и компьютерным зрением.	Сложность настройки и использования без технической подготовки, отсутствие инструментов для точной медицинской аннотации.
SlideRunner [9]	Инструмент с открытым исходным кодом, предназначенный для аннотирования цифровых слайдов. Он предоставляет интерфейс для ручной разметки, а также интеграцию с библиотеками машинного обучения.	Отсутствие удобного интерфейса для работы с изображениями сверхбольшого размера, недостаточная оптимизация для аннотирования сложных структур, ограниченные возможности автоматизации процесса разметки.
Облачные инструменты: AWS SageMaker, Google Cloud AI	Предлагают интеграцию с облачными вычислениями и поддерживают работу с большими данными.	Высокая стоимость использования, зависимость от интернет-соединения, отсутствие специализированных функций для работы с медицинскими изображениями.

Описание алгоритма бинаризации и выделения контуров

Визуализация результатов каждого этапа обработки изображений представлены на рис. 2.

Процесс бинаризации цветного изображения (рис. 2, а) выполняется с использованием алгоритмов обработки изображений, предоставляемых библиотекой **OpenCV**, и включает несколько этапов

Цветное изображение в формате **OpenCV** имеет три цветных канала в формате **BGR**. В результате проведенных экспериментов было установлено, что бинаризацию наиболее эффективно проводить по красному каналу (**channel** = 2).

Для выделения объектов на изображении используется метод пороговой бинаризации. Результат бинаризации — двоичное изображение **binary_image**, где объекты выделены белым цветом на чёрном фоне. Для улучшения качества бинарного изображения применяется морфологическая обработка, которая помогает устранить шумы и выделить структуры на изображении. Изображение, полученное в результате бинаризации, представлено на рис 2, б.

Алгоритм выделения контуров состоит из нескольких шагов:

- выделение начальных прямоугольников (рис. 2, в);

- объединение пересекающихся прямоугольников;

- объединение близко расположенных прямоугольников;

- фильтрация прямоугольников;

- повторное объединение пересекающихся прямоугольников.

Этот алгоритм используется для выделения областей интереса на изображении. Он позволяет уменьшить количество ложных срабатываний, объединить пересекающиеся и близкие области, а также сохранить только значимые регионы (рис. 2, г).

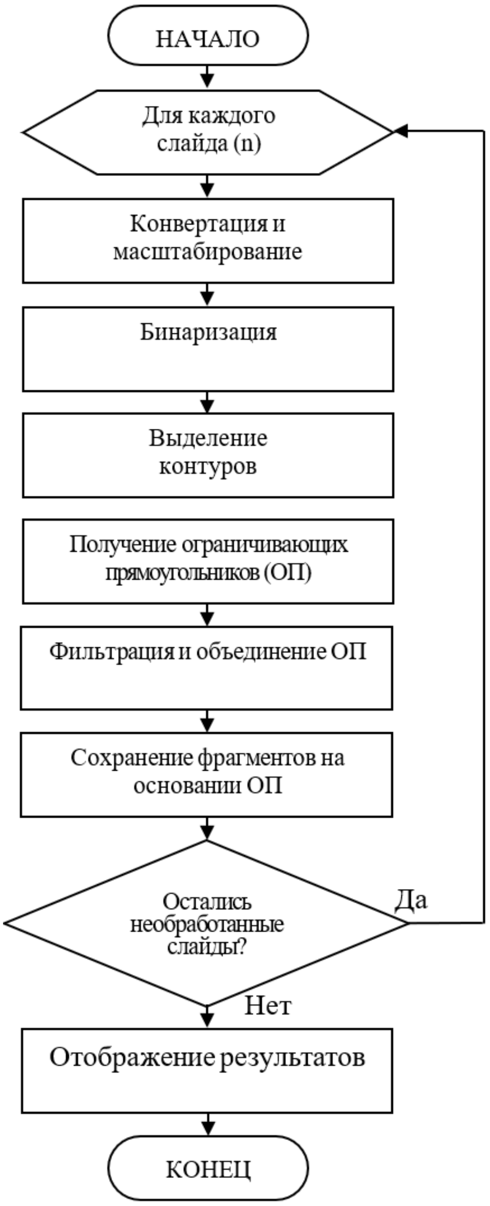

Рисунок 1 – Обобщенный алгоритм обработки изображения

Подготовленные фрагменты (рис. 2, д) сохраняются в отдельные файлы, в именах которых закодирована информация, необходимая для последующего переноса аннотаций на оригинальное изображение высокого разрешения. Эта информация включает имя исходного файла, координаты фрагмента относительно исходного изображения, ширину и высоту фрагмента до масштабирования.

Для разметки подготовленных фрагментов используется **Makesense.AI** [10] – веб-инструмент, который отличается простым и интуитивно понятным интерфейсом, поддержкой различных форматов аннотации (например, **JSON, XML**), а также гибкостью в настройке классов и типов аннотаций, что важно для патологоанатомов, работающих с уникальными структурами биопсийных материалов.

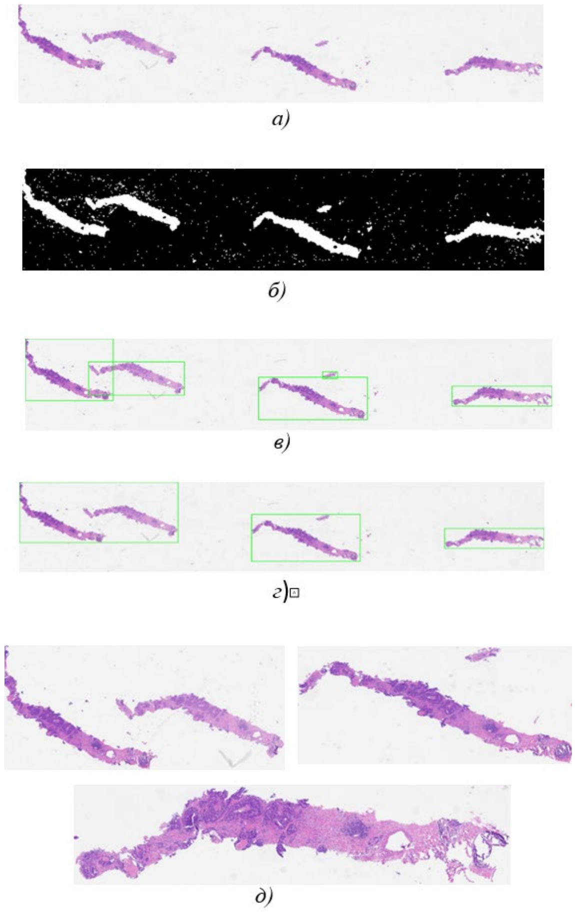

*Рисунок 2 – Этапы подготовки изображения: а) изображение сконвертировано в формат **JPEG** выбранного разрешения; б) произведен процесс бинаризации; в) выделены первичные ограничивающие прямоугольники для найденных объектов; г) близкие и пересекающиеся ограничивающие прямоугольники объединены; д) сохранены отдельные фрагменты изображения*

Заключение

В данной статье были рассмотрены основные сложности аннотирования медицинских изображений большого размера, таких как снимки биопсии, и предложено комплексное решение для их преодоления.

Предложенная методика использует библиотеку **large-image** с **tifffile** в качестве бэкэнда для обработки **TIFF**-изображений. Конвертация исходных данных в формат **JPEG** с необходимым разрешением позволяет значительно уменьшить размер файлов при сохранении качества, необходимого для аннотирования. Методы бинаризации и поиска контуров автоматизируют выделение областей интереса, а разбиение изображения на фрагменты упрощает процесс разметки. Сохранение метаданных, таких как координаты вырезанных областей, в именах файлов позволяет отслеживать связь между исходными и аннотированными данными.

Для аннотирования полученных фрагментов был выбран инструмент **Makesense.AI** благодаря его простоте, гибкости и доступности для специалистов без технической подготовки. Такой подход обеспечивает удобство работы для патологоанатомов и минимизирует временные затраты на разметку.

Литература

1. Litjens, G., Kooi, T., Bejnordi, B. E., et al. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017, 42: 60–88. DOI: 10.1016/j.media.2017.07.005

2. Ronneberger, O., Fischer, P., Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, 9351: 234–241. DOI: 10.1007/978-3-319-24574-4_28

3. Aresta, G., Araújo, T., Kwok, S., et al. BACH: Grand challenge on breast cancer histology images. Medical Image Analysis, 2019, 56: 122–139. DOI: https://doi.org/10.1016/j.media.2019.05.010.

4. Large-Image Documentation. Kitware, Inc. https://doi.org/10.5281/zenodo.14624225 Интернет-ресурс. - Режим доступа : www/ URL: https://girder.github.io/large_image/

5. OpenCV Documentation. Open Source Computer Vision Library. Интернет-ресурс. - Режим доступа : www/ URL: https://docs.opencv.org/

6. Bankhead, P., Loughrey, M. B., Fernández, J. A., et al. QuPath: Open source software for digital pathology image analysis. Scientific Reports, 2017,7(1): 16878. DOI: 10.1038/s41598-017-17204-5.

7. Rueden, C. T., Schindelin, J., Hiner, M. C., et al. ImageJ2: ImageJ for the next generation of scientific image data. BMC Bioinformatics, 2017,18(1): 529. DOI: 10.1186/s12859-017-1934-z.

8. Schindelin, J., Arganda-Carreras, I., Frise, E., et al. Fiji: An open-source platform for biological-image analysis. *Nature Methods*, 2012, 9(7): 676–682. DOI: 10.1038/nmeth.2019

9. M. Aubreville, C. Bertram, R. Klopfleisch and A. Maier (2018) SlideRunner - A Tool for Massive Cell Annotations in Whole Slide Images. In: Bildverarbeitung für die Medizin 2018. Springer Vieweg, Berlin, Heidelberg, 2018. pp. 309-314. https://doi.org/10.1007/978-3-662-56537-7_81 Режим доступа : www/ URL: https://github.com/maubreville/SlideRunner

10. Makesense.AI. Онлайн-инструмент для аннотирования изображений. Интернет-ресурс. - Режим доступа : www/ URL: https://www.makesense.ai/

Хрюкин Е.А., Мартыненко Т.В., Васяева Т.А., Швороб Д.С. Разработка комплексной методики для аннотирования изображений биопсии. *В статье предложена комплексная методика аннотирования медицинских изображений биопсии на основе языка **Python** и библиотеки **large-image**, включающая конвертацию изображений в формат **JPEG**, автоматическое выделение участков биопсии, их разбиение на фрагменты и использование инструмента **Makesense.AI** для аннотирования, позволяющая сэкономить вычислительные ресурсы и упростить работу специалистов, обеспечивающая эффективную подготовку данных для обучения нейросетей.*

Ключевые слова: *пороговая бинаризация, выделение контуров, large-image, tifffile, Python.*

Yevgeniy A. Khriukin, Tatyana V. Martynenko, Tatyana A. Vasyaeva, Danil S. Shvorob. Development of a comprehensive method for biopsy image annotation. *This paper presents a comprehensive approach to annotating medical biopsy images using **Python** and the **large-image** library. The approach includes converting images to **JPEG** format, automatically detecting biopsy regions, dividing them into fragments, and employing the **Makesense.AI** tool for annotation. This method helps save computational resources and simplifies the work of specialists, making it an effective tool for preparing data for neural network training.*

Keywords: *threshold binarization, contour detection, large-image, tifffile, Python.*

Статья поступила в редакцию 02.03.2025
Рекомендована к публикации профессором Скобцовым Ю. А.