Паспорт набора данных

1. Идентификатор набора данных

DS052023

2. Версия набора данных

1.0

3. Наименование набора данных

«300 клинических фотографий зубных рядов взрослых пациентов во фронтальной и боковых проекциях для обучения нейросети автоматической сегментации (разметка полигонами)»

4. Описание набора данных.

Аннотация

Одним из направлений стратегического развития России является разработка цифровых технологий и создание высокотехнологичных сервисов. В том числе предполагается разработка новых сервисов, связанных с компьютерным зрением, обработкой естественного языка, системами поддержки принятия решений, автоматизацией обучения нейронных сетей и др. (Бузова, 2020). Использование искусственного интеллекта для решения различных задач позволяет автоматизировать рутинные, трудоемкие или сложные процессы, а также повысить их точность и производительность (Ossowska, 2022).

Компьютерный анализ медицинских изображений является перспективным направлением, которое позволит автоматизировать и объективизировать процесс описания и оценки данных изображений. Медицинские изображения (рентгенограммы, магнитно-резонансные томограммы, компьютерные томограммы, клинические фотографии и др.) играют важную роль в диагностике, лечении и профилактике заболеваний (Zhu, 2020). В стоматологии клинические фотографии пациентов являются особенно важной частью медицинской документации. Такие изображения могут помочь в

диагностическом процессе путем документирования изменений, происходящих с возрастом или в процессе лечения, получения экспертного мнения от других практикующих врачей-стоматологов или заключения искусственного интеллекта, реализованного в системе поддержки принятия врачебных решений (Hood, 1998).

Одним из первых этапов создания систем поддержки принятия врачебных решений является автоматическая сегментация изображений. Сегментация изображения — это процесс разделения изображения на элементы со схожими характеристиками и свойствами (Majanga, 2021). Сегментация изображения зубных рядов с помощью машинного обучения является сложной задачей из-за высокой изменчивости формы зубов. В связи с этим необходима разработка обученной сети и модели машинного обучения, способной сегментировать зубы с высокой точностью (El Bsat, 2022).

На сегодняшний день количество клинических фотографий зубных рядов, официально доступных для тренировки систем глубокого обучения в исследовательских и образовательных целях, ограничено (Majanga, 2021), поэтому вопрос создания базы данных подобных клинических фотографий является особенно актуальным.

Клиническая задача

Автоматическая сегментация изображений зубов на клинических фотографиях для последующего их использования в создании систем поддержки принятия решений для диагностики состояния твердых тканей зубов

Назначение набора данных

Набор размеченных клинических фотографий можно будет использовать для обучения нейросетей осуществлению автоматической сегментации изображений зубов. Сегментация является одним из первых этапов создания систем поддержки принятия врачебных решений по типу «Компьютерного зрения»

Нозологии

К08.1 Потеря зубов вследствие несчастного случая, удаления или локальной периодонтальной болезни

5. Владелец набора данных

ФГАОУ ВО Первый МГМУ им И.М. Сеченова

Контактные данные

119991, г. Москва, ул. Трубецкая, д. 8, стр. 2

Авторы

Макеева И.М., Новожилова Н.Е., Бабина К.С., Полякова М.А.

6. Порядок предоставления доступа к набору данных

Будет уточнен

7. Тэги

ДАТАСЕТ, НЕЙРОННЫЕ СЕТИ, РАЗМЕТКА ДАННЫХ, КЛИНИЧЕСКИЕ ФОТОГРАФИИ ЗУБОВ, ОТСУТСТВИЕ ЗУБОВ, СЕГМЕНТАЦИЯ ИЗОБРАЖЕНИЙ

8. Источник данных

Глубина набора данных

06.2023–12.2024

Объем набора данных

300 клинических фотографий (общий объем будет уточнен)

Критерии отбора клинических параметров и пациентов и соответствующих, а также виды лабораторно-инструментальных исследований

Взрослые (> 18 лет) пациенты; подписанное информированное согласие; наличие хотя бы одного зуба; наличие клинических фотографий удовлетворительного качества.

Особенности подготовки исследований

Обрезка фотографий (с целью выделения области интереса)

Половозрастные характеристики набора данных

Пациенты обоих полов старше 18 лет

Характеристики оборудования

Фотоаппарат с разрешением матрицы от 10 Мегапикселей, светочувствительностью не менее 200 ISO, глубиной цвета не менее 10 бит; с объективом с минимальной апертурой f/1,8, фокусным расстоянием не менее 20 мм; с возможностью записи изображений без обработки (в формате raw).

Дата публикации набора данных

До конца 2024 года

Дата обновления набора данных

Не обновлялся

9. Модели машинного обучения

Mask R-CNN

10. Расположение набора данных

Единая база данных Сеченовского Университета

11. Структура набора данных

Схема структуры директорий и файлов

Набор данных хранится в формате COCO (Common Objects in COntext)

<dataset_dir>/

data/

<filename0>.<ext>

...

labels.json

Где dataset_dir – корневой каталог набора данных, data – подкаталог, содержащий файлы с изображениями (<filename0>.<ext> и тд), labels.json – структура, описывающая метки и метаданные (в том числе возраст, пол, и тд) разметки для набора данных в формате COCO.

Описание структуры и принципа наименования файлов и директорий.

Имя файла формируется по принципу UUID (universally unique identifier «всемирно уникальный идентификатор») стандарт идентификации, используемый в создании программного обеспечения, стандартизированный Open Software Foundation (OSF) как часть DCE — среды распределённых вычислений. Основное назначение UUID — это позволить распределённым системам уникально идентифицировать информацию без центра координации.

12. Обзор разметки данных

Особенности формата разметки

Сегментация экземпляров (instance segmentation)

Классы разметки

Зуб, тип зуба, коронка

Принципы разметки

Многоугольником (polygon)

Принципы верификации разметки

Разметку выполняет один специалист, второй проверяет и подтверждает корректность размеченных данных

Статистика использования лейблов и классов

Будет уточнена

13. Правила использования и распространения

Копирайт

ФГАОУ ВО Первый МГМУ им И.М. Сеченова