Технологии лучевой диагностики | Radiological technologies

ISSN 1607-0763 (Print); ISSN 2408-9516 (Online)
https://doi.org/10.24835/1607-0763-1263

Распознавание областей текста с персональными данными на диагностических изображениях

© **Новик В.П.[1], Кульберг Н.С.[2], Арзамасов К.М.[1], Четвериков С.Ф.[1], Хоружая А.Н.[1]*, Козлов Д.В.[1], Кремнева Е.И.[1, 3]**

[1] ГБУЗ "Научно-практический клинический центр диагностики и телемедицинских технологий ДЗ города Москвы"; 109029 Москва, Средняя Калитниковская ул., д. 28, стр. 1, Российская Федерация

[2] ФГУ "Федеральный исследовательский центр "Информатика и управление" Российской академии наук"; 119333 Москва, ул. Вавилова, д. 44, корп. 2, Российская Федерация

[3] ФГБНУ "Научный центр неврологии"; 125367 Москва, Волоколамское шоссе, д. 80, стр. 1, Российская Федерация

Цель исследования: разработка метода обнаружения областей текста с приватными данными на медицинских диагностических изображениях при помощи модуля Tesseract и модифицированного расстояния Левенштейна.

Материал и методы. Для пороговой фильтрации на начальном этапе определяется яркость точек, принадлежащих символам текста на изображениях. Динамический порог вычисляется по гистограмме яркостей пикселей изображения. Далее для первичного распознавания текста используется модуль Tesseract. На основании значений тэгов из DICOM-файлов формировался набор строк для поиска их в распознанном тексте. Для поиска этих строк использовалось модифицированное расстояние Левенштейна. Для тестирования алгоритма применялся набор DICOM файлов типа "Dose Report" модальности CT. Оценку точности проводили эксперты, размечающие блоки приватной информации на изображениях.

Результаты. Разработан инструмент с набором метрик и оптимальных порогов для выбора решающих правил в нахождении совпадений, позволяющих обнаруживать области текста с приватными данными на медицинских изображениях. Для этого инструмента определена точность локализации областей с личными данными по сравнению с экспертной разметкой, которая составляет 99,86%.

Заключение. Разработанный в рамках настоящего исследования инструмент позволяет выявлять персональные данные на цифровых медицинских изображениях с высокой точностью, что указывает на возможность его практического применения при подготовке наборов данных.

Ключевые слова: оптическое распознавание текста, медицинские изображения, свёрточные нейронные сети, Tesseract, анонимизация данных, DICOM

Авторы подтверждают отсутствие конфликтов интересов.

Для цитирования: Новик В.П., Кульберг Н.С., Арзамасов К.М., Четвериков С.Ф., Хоружая А.Н., Козлов Д.В., Кремнева Е.И. Распознавание областей текста с персональными данными на диагностических изображениях. *Медицинская визуализация.* 2023. https://doi.org/10.24835/1607-0763-1263

Поступила в редакцию: 07.10.2022. **Принята к печати:** 14.12.2022. **Опубликована online:** 17.07.2023.

Recognition of text areas with personal data on diagnostic images

© **Vladimir P. Novik[1], Nicholas S. Kulberg[2], Kirill M. Arzamasov[1], Sergey F. Chetverikov[1], Anna N. Khoruzhaya[1]*, Dmitriy V. Kozlov[1], Elena I. Kremneva[1, 3]**

[1] Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of Moscow Health Care Department; 24, Petrovka str., Moscow 127051, Russian Federation

[2] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences; 44, Vavilova str., Moscow 125367, Russian Federation

[3] Research center of neurology; 80-1, Volokolamskoye shosse, Moscow 125367, Russian Federation

The aim of the study is to develop a method for detecting areas of text with private data on medical diagnostic images using the Tesseract module and the modified Levenshtein distance.

Materials and methods. For threshold filtering, the brightness of the points belonging to the text characters in the images is determined at the initial stage. The dynamic threshold is calculated from the histogram of the brightness of the pixels of the image. Next, the Tesseract module is used for primary text recognition. Based on the tag values from DICOM files, a set of strings was formed to search for them in the recognized text. A modified Levenshtein distance was used to search for these strings. A set of DICOM files of the "Dose Report" type was used to test the algorithm. The accuracy was assessed by experts marking up blocks of private information on images.

Results. A tool has been developed with a set of metrics and optimal thresholds for choosing decisive rules in finding matches that allow detecting areas of text with private data on medical images. For this tool, the accuracy of localization of areas with personal data on a set of 1131 medical images was determined in comparison with expert markup, which is 99.86%.

Conclusion. The tool developed within the framework of this study allows identifying personal data on digital medical images with high accuracy, which indicates the possibility of its practical application in the preparation of data sets.

Keywords: optical character recognition (OCR), medical images, convolutional neural networks, Tesseract, DICOM, data anonymization

Conflict of interest. The authors declare no conflict of interest. The study had no sponsorship.

For citation: Novik V.P., Kulberg N.S., Arzamasov K.M., Chetverikov S.F., Khoruzhaya A.N., Kozlov D.V., Kremneva E.I. Recognition of text areas with personal data on diagnostic images. *Medical Visualization.* 2023. https://doi.org/10.24835/1607-0763-1263

Received: 07.10.2022. **Accepted for publication:** 14.12.2022. **Published online:** 17.07.2023.

Введение

В современной медицинской практике проводится огромное количество радиологических исследований. В дальнейшем многие такие исследования включаются в состав специализированных наборов данных, которые передаются для сбора статистики в контролирующие организации, используются в научных исследованиях, при обучении систем искусственного интеллекта и т.п. Несмотря на то что при формировании наборов данных исследования анонимизируются, остается риск раскрытия приватной информации о пациентах и медицинском персонале. И в этих условиях проблема защиты персональных данных встает очень остро.

Данные о пациентах в радиологических изображениях хранятся в специальных структурах DICOM-файлов [1]. Стандартные методы анонимизации реализуются с помощью замены или уничтожения тэгов приватной информации [2]. Однако часть персональных данных может быть закодирована непосредственно в пикселях изображений (например, в дозовых отчетах, вторичных объемных реконструкциях и др.). Для этих целей специально создан тэг с "прожигаемыми" персональными данными, но он не всегда заполняется, и данные могут оставаться. Удаление приватной информации, внедренной в изображение, опирается на методы оптического распознавания текста (Optical character recognition, OCR). Для этого необходимо найти и маскировать соответствующие участки изображения.

Судя по данным литературы, лидирующими направлениями разработок программного обеспечения для анализа медицинских изображений, в обучении которых требуются наборы диагностических исследований, являются алгоритмы для выявления патологий органов грудной клетки и центральной нервной системы при компьютерной томографии (КТ) [3]. Поэтому в данной работе исследовались изображения модальности CT "Dose report" ("Дозовый отчет") на устройствах Toshiba, наиболее распространенных в медицинских учреждениях России (по данным Формы №30). Данные типы изображений генерируются программно-аппаратным комплексом на основе первичных данных о пациенте. Каждый такой файл хранит набор тэгов, общий для всего исследования в модальности КТ. Но, кроме того, эти файлы содержат изображения с текстом. Нашей задачей является выделение на этих изображениях областей, содержащих текст с персональными данными. Личной информацией являются имена пациентов или медицинского персонала, дата рождения, пол, возраст, идентификаторы исследования.

Распознавание текста с персональными данными на медицинских изображениях – обширная тема. В работах [4, 5] для OCR используют фиксированный набор шаблонов символов. Этот подход применим, когда заранее известны глифы всех используемых символов. Он обеспечивает гарантированно высокую точность распознавания до 97%. Следует отметить, что для тестирования точности алгоритмов в разных работах использу-

ются разные базы данных и разнообразные метрики. Был разработан [6] интерактивно-адаптационный алгоритм, идентифицирующий 3 класса личной информации, обученный на 15 334 изображениях. Вероятность ложноположительного обнаружения составляет менее 4%.

Авторы другой статьи [7] разработали метод выделения личной информации на ультразвуковых изображениях, использующий свёрточные нейронные сети. Точность составляет 89,2% на 500 медицинских изображениях. Наиболее близкой по тематике нашей работы нужно отметить статью G. Kin и соавт. [8]. Ее авторы выделяли текст с личными данными в изображениях нескольких модальностей и получили 0% вероятности ложного обнаружения и 0,5% вероятность пропуска на 660 синтезированных изображениях. Однако в данной статье не указаны метрики, использованные при поиске совпадения распознанного текста и строк из DICOM-тэгов. Также не описаны критерии принятия решения при нахождении совпадения.

Цель исследования: оценка применимости метода обнаружения областей текста с приватными данными на реальных медицинских диагностических изображениях.

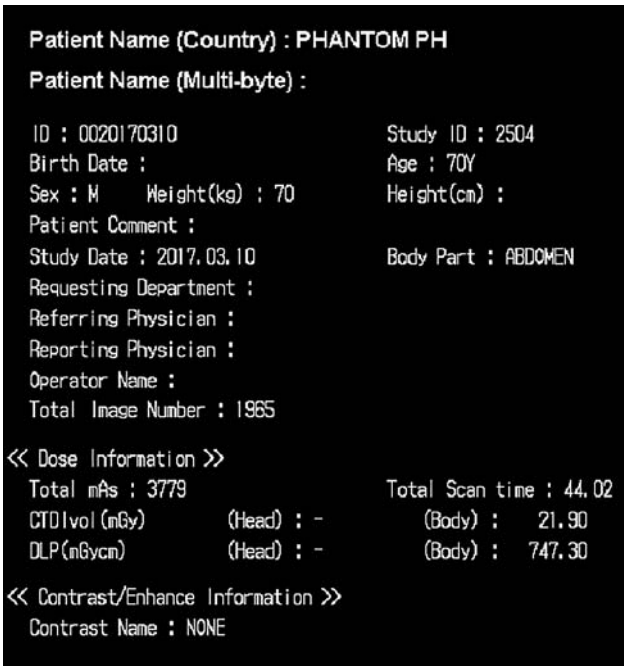

Рис. 1. Пример изображения класса 'dosereport'. Белые символы на черном фоне.

Fig. 1. Example of an image of the 'dosereport' class. White symbols on a black background.

Материал и методы

Обрабатываемые данные

Сформирован набор медицинских данных из 1131 медицинского изображения в формате DICOM. На рис. 1 представлен пример такого изображения ***I*** (дозовый отчет).

Данный файл получен при сканировании искусственного фантома на программно-аппаратном комплексе модели "Aquilion". Поскольку обследовался фантом, данные в этом файле не являются приватными и используются для иллюстрации работы алгоритма. Изображение ***I*** представляет собой двухмерную матрицу размером 512 × 512 чисел. Значения яркостей лежат в диапазоне (–32768,32768). Фону соответствует низкое значение яркости. Текст в изображении ***I*** состоит из букв латинского алфавита, цифр, скобок и знаков препинания. Используется два шрифта с символами верхнего и нижнего регистра. Используемые шрифты, положение блоков информации могут изменяться в зависимости от типа программно-аппаратного комплекса, на котором создаются файлы. Помимо изображения, файл содержит персональную информацию в DICOM-тэгах ***T***:

$$\boldsymbol{T} = \{tag_i: Value(tag_i), \forall\, tag_i \in (tag\_to\_search)\},$$

Value( ) – функция извлечения значений, соответствующих этому тэгу из медицинского файла. Нас будет интересовать следующий набор тэгов *{tag_to_search}*, имеющих отношение к личной информации:

- *PatientName* – имя пациента,
- *PatientID* – идентификатор пациента в базе данных,
- *PatientAge* – возраст пациента,
- *PatientSex* – пол пациента,
- *PatientBirthDate* – дата рождения пациента,
- *StudyID* – идентификатор проведенного исследования,
- *AccessionNumber* – еще один идентификатор проведенного исследования в базе данных.

Каждому тэгу в медицинском файле может соответствовать одно или несколько значений. Например, тэгу '*PatientID*' может соответствовать несколько значений, если пациенту назначались разные идентификаторы для хранения в разных системах баз данных. Какое именно из значений будет представлено на изображении, заранее неизвестно. Для файла "PHANTOM.dcm" тэгам с личными данными соответствуют следующие значения:

Value(PatientName) = 'ФАНТОМ Ф';
Value(PatientID) = 0020170310;
Value(PatientAge) = 70Y;
Value(PatientSex) = M;
Value(StudyID) = 2504.

Как правило, имя пациента хранится в национальном алфавите, в России на русском. А в изображении внедряется на английском языке. Поэтому значение данного тэга транслитерируется в латинский алфавит:

$$Value(PatientName) = PHANTOM\ PH.$$

Если тэг с приватной информацией не заполнен, то данный тэг исключается из дальнейшего поиска на изображении. Например, для изображения на рис. 1 был исключен тэг *PatientBirthDate*.

Формирование строк для поиска

Тэгу '*PatientSex*' соответствует значение, состоящее всего из одного символа. Очевидно, что поиск на изображении текста из одного символа приведет к большому количеству ложных обнаружений. Значения тэгов 'AccessionNumber' = '2504', 'StudyID' = '2504' состоят всего из 4 символов. Поиск коротких строк опасен ложными совпадениями. Для уменьшения вероятности ложного совпадения к строке поиска добавляется значение префикса, который обычно предшествует данному тэгу на этом изображении:

- 'StudyID' – 'Study ID:',
- 'AccessionNumber' – 'Accession No:',
- 'PatientSex' – 'Sex:',
- 'PatientAge' – 'Age:',
- 'PatientName' – '' (не используется),
- 'PatientID' – 'ID',
- 'PatientBirthDate' – 'Birth Date',
- 'SeriesNumber' – 'Number'.

Если будут обрабатываться медицинские изображения, для которых характерны другие префиксы, то формируется другой набор префиксов. Для тэга 'PatientName' префикс не используется, так как имя пациента, как правило, состоит из достаточного числа символов для надежного нахождения его в распознанном тексте. Таким образом, из значений тэгов будут сформированы наборы строк для поиска в распознанном тексте:

$$\boldsymbol{S_{etalon}} = \{S_{etalon}(k)\} = \{prefix(tag_k) + T(tag_k)\},$$
$$tag_i \in (tag\_to\_search)\},$$

где *prefix()* – функция формирования префикса, соответствующего этому тэгу для данного типа изображений.

Предобработка входного изображения

Точки, принадлежащие символам текста на изображениях ***I***, имеют высокие значения яркости. На некоторых изображениях с текстом могут присутствовать также изображения графиков и предварительного просмотра обследуемых органов. Для маскирования сторонних объектов применяются пороговая фильтрация и удаление крупных связанных объектов.

Пороговая фильтрация. Вычисляется динамический порог по гистограмме яркостей пикселей изображения:

$$THR\_BR = max\{I(x,y)\},(x,y) \in \boldsymbol{I}.$$

Этот порог используется для бинарной классификации пикселей изображения:

$$\boldsymbol{I\_thr} = \{I(x,y) \geqslant THR\_BR,\}(x,y) \in \boldsymbol{I}.$$

Формирование связанных компонент точек:

$$\boldsymbol{I\_label} = label(\boldsymbol{I\_thr}).$$

Удаление кластеров с линейными размерами, существенно превышающими размер символа, с формированием изображения ***I_final***. Используется фиксированный порог фильтрации, подобранный эмпирическим способом THR_label = 40. Данный порог показал удовлетворительное качество фильтрации остаточных шумовых объектов.

Распознавание текста модулем Tesseract

Предобработанное изображение ***I_final*** обрабатывается алгоритмом Tesseract 4.1.3 [9]. Вероятность правильного распознавания символа на используемых изображениях составляет около 80–90%. Особую сложность для Tesseract представляют алфавитно-цифровые последовательности. А именно такие слова часто встречаются для тэгов 'AccessionNumber', 'StudyID', 'PatientID'. Средняя точность распознавания символов в таких словах падает до 65%. Тем не менее качество работы модуля Tesseract является одним из лучших среди альтернативных библиотек с открытым кодом. Мы не рассматриваем возможность обучения (переобучения) модуля Tesseract, так как это требует априорных знаний о каждом символе всех шрифтов, использующихся в обрабатываемых изображениях, которые отсутствуют. При гипотетическом наличии такой информации можно проводить OCR корреляционным методом поиска шаблонов.

Результатом работы модуля Tesseract является набор из *N_words* слов, распознанных слов на изображении. Для каждого распознанного слова выдается набор характеристик, из которых существенными для нашей задачи являются:

$$text\_box_i = ['left', 'top', 'width', 'height'] –$$

координаты прямоугольника на изображении, в котором находится распознанное *i*-е слово, $^0text^0$ – текст распознанного слова.

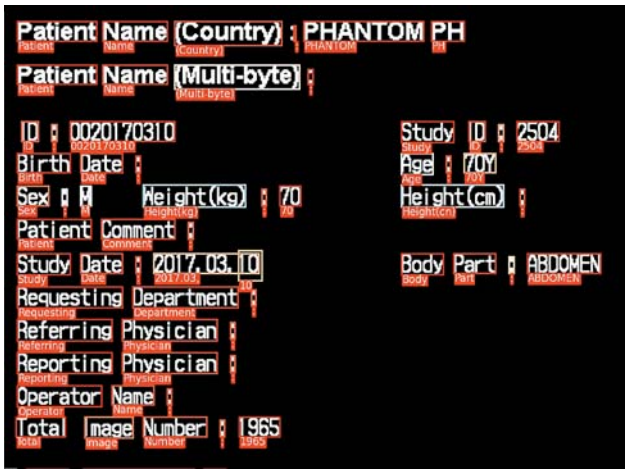

Рис. 2. Результаты распознавания библиотекой Tesseract.

Fig. 2. The results of recognition by the Tesseract library.

Таким образом, результатом распознавания текста является набор данных ***S***:

$$S = \{[text_i, text\_box_i]\},\ i \in [0, N\_words).$$

Для файла Фантом результаты визуализации распознавания представлены на рис. 2. Каждое распознанное слово заключено в прямоугольник $text\_box_i$, снизу которого напечатано соответствующее значение '$text_i$'.

Поиск соответствий значений DICOM-тэгов и распознанных символов

На данном этапе мы имеем результаты распознавания текста ***S*** и список эталонных строковых переменных для поиска $\boldsymbol{S_{etalon}}$. Задача, выполняемая в данном блоке, состоит в нахождении наилучшего соответствия для каждой переменной $\boldsymbol{S_{etalon}(k)}$ в результатах распознавания текста ***S***. Для поиска соответствий используется расстояние Левенштейна [10, 11], нормированное на длину максимальной строки. На рис. 2 можно видеть пример, что слово “2017.03.10” расщеплено на 2 слова “2017.03” и “10”. Одному слову на изображении могут соответствовать несколько слов в распознанном тексте. При поиске некоторой строки в распознанном тексте неизвестно заранее, будет ли эта строка соответствовать одному слову или нескольким словам. Поэтому из распознанного текста ***S*** формируются все возможные строки длиной до *maxlength* слов. Для каждого индекса $i \in [0, N\_words)$ формируется набор строковых переменных ***W*** путем объединения слов из ***S*** с индексами [1 : *maxlength*]:

$$\boldsymbol{W} = \{W(i,j)\},$$

где $W(i,j)$ является слиянием в одну строковую переменную слов

$$[text_i, text_{i+1}, \ldots, text_{i+j-1}].$$

Каждый элемент $W(i,j)$ сравнивается с эталонными словами из $\{S_{etalon}(k)\}$, формируя матрицу расстояний ***D***:

$$\boldsymbol{D} = D(i, j, k) = \frac{dist(W(i, j), S_{etalon}(k))}{\max(len(W(i,j)) + len(S_{etalon}(k)))},$$

где *dist()* – функция вычисления расстояния Левенштейна, *len()* – длина слова.

Эмпирически на нашем наборе данных выбрано значение *maxlength* = 7, которое гарантированно обеспечивает формирование слов нужной длины на собранной базе изображений. **Определение непересекающихся соответствий с минимальными расстояниями.** Некоторые тэги могут присутствовать на изображении не один, а несколько раз N_{cases}. Поэтому для каждого тэга *k* из множества сравнений ***D*** выбирается N_{cases} минимальных элемента *min_dist_tag(k)*, расположенных в непересекающихся прямоугольниках. При выборе трех минимальных случаев также вводится дополнительное ограничение: $D_i \geqslant D_0 + 0.1$. Это условие позволяет устранить случаи значительных расстояний при поиске одного и того же тэга. Выбор числа 3 обусловлен максимальным возможным числом текста тэга в изображении. Например, в некоторых изображениях имя пациента может быть представлено 3 раза. И расстояние ***D*** для этих трех случаев не может значительно отличаться, так как сравниваются примерно одни и те же данные.

Мы сортируем все расстояния для тэга $S_{etalon}(k)$ по возрастанию, наиболее похожие будут на первых местах. Выбираем первое расстояние и соответствующий ему прямоугольник $text\_box_i$ и добавляем их к списку *min_dist_tag(k)*. Для каждого следующего расстояния проверяем, нет ли пересечений с уже добавленными прямоугольниками в *min_dist_tag(k)*. Если нет, то добавляем его в *min_dist_tag(k)*. Таким образом мы получаем для каждого тэга k список $min\_dist\_tag(k) = \{[dist_i, rect_i]\}$. На рис. 3 представлен результат поиска соответствий. Для каждой области, изображенной красным прямоугольником, указаны тип тэга и расстояние соответствия (снизу прямоугольника).

Следующим шагом являются проверка и устранение конфликтов пересечений блоков разных тэгов. Это решается похожим методом. Информация о потенциальных соответствиях отдельных блоков *min_dist_tag(k)* объединяется, сортируется и фильтруются случаи пересечения блоков. При конфликте пересечений выбирается блок с мини-

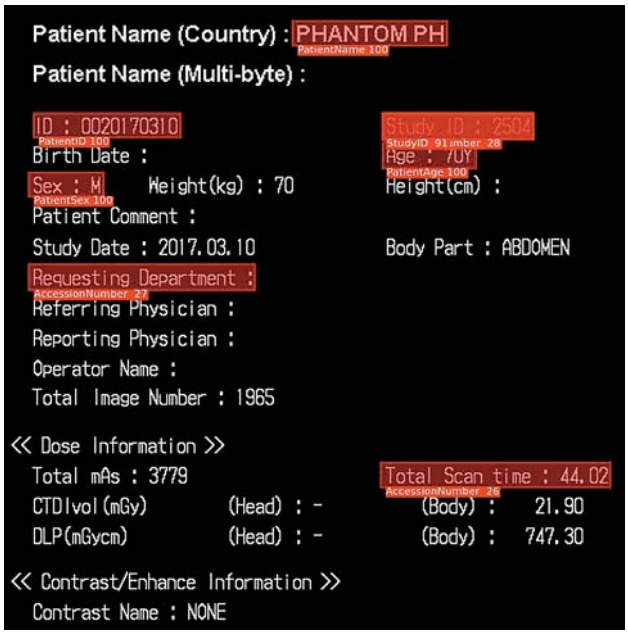

Рис. 3. Потенциальные области локализации приватной информации (красные прямоугольники).

Fig. 3. Potential areas of localization of private information (red rectangles).

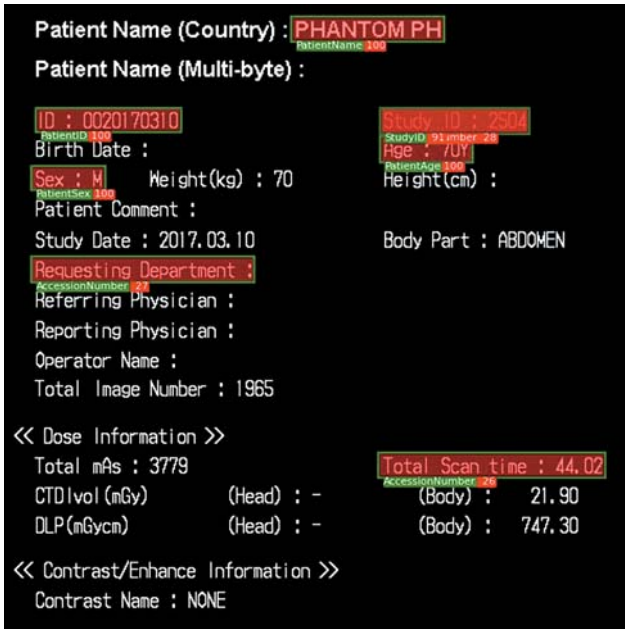

Рис. 4. Потенциальные области локализации приватной информации (зеленые прямоугольники) после фильтраций пересечений всех тэгов.

Fig. 4. Potential areas of localization of private information (green rectangles) after filtering the intersections of all tags.

мальным расстоянием. Это дает нам конечный результат об обнаруженных соответствиях эталонных строк и распознанного текста *min_dist_tag(k)*. Для каждого обнаруженного прямоугольника имеется расстояние соответствия. На рис. 4 представлен окончательный результат поиска соответствий для всех тэгов личной информации. Для каждой области, изображенной зеленым прямоугольником, указаны тип тэга и расстояние соответствия (снизу прямоугольника).

Результаты

База данных для тестирования алгоритма

Для тестирования алгоритма использовался набор из 1131 DICOM-файла “Dose Report” модальности CT. Для оценки точности использовалась экспертная разметка блоков приватной информации на изображениях. Для каждого изображения эксперты разметили области с приватной информацией. Каждая область представляет собой прямоугольник, для которого отмечено, какому тэгу он принадлежит. Разметка проводилась с помощью программного обеспечения [12].

Программная реализация алгоритма

Описанный метод был реализован в виде программного модуля на языке Python. Работа с DICOM-файлом (чтение тэгов, загрузка изображения) реализована через библиотеку pydicom 2.2.1 [13]. Предобработка выполняется функциями из библиотеки opencv 4.4.0 [14].

Время работы алгоритма

Среднее время обработки одного файла изображения составляет 0,9 с. Операции чтения тэгов и загрузка изображения занимают 0,2 с. Распознавание текста модулем Tesseract требует 0,4 с на одно изображение. Поиск соответствий значений DICOM-тэгов и распознанных символов занимает 0,3 с. Для тестирования времени выполнения алгоритма использовался персональный компьютер с процессором Core(TM) i5-8400 и 16 Гб оперативной памяти.

Выбор оптимального решающего порога

Для каждого детектированного прямоугольника наш алгоритм определяет расстояние *Dist* между текстом в этом прямоугольнике и эталонной строкой. Чем больше это расстояние, тем больше вероятность ложного обнаружения прямоугольника. Используем решающий порог *Dist_thresh*, определяющий итоговое решение о наличии детектированной области: если *Dist* < *Dist_thresh*, то область соответствует приватному тэгу.

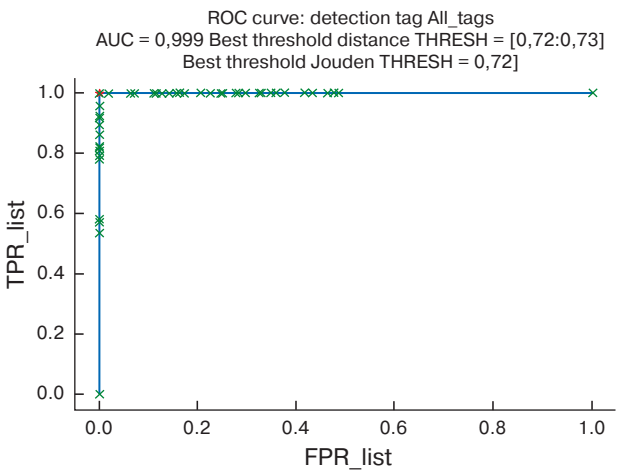

Рис. 5. *ROC*-кривая локализации областей приватной информации.

Fig. 5. *ROC*-curve of localization of private information areas.

Таблица. Число ошибок локализации областей с приватной информацией

Table. The number of localization errors in areas with private information

Название тэга Name of tag	ИП TP	ИО TN	ЛП FP	ЛО FN
PatientSex	639	492	0	0
PatientName	628	503	0	0
PatientAge	636	492	0	3
AccessionNumber	575	552	0	4
StudyID	663	468	0	0
PatientID	446	681	1	3
PatientBirthDate	639	492	0	0
Все тэги / All tags	4226	3680	1	10

Примечание. ИП – истинно положительный, ИО – истинно отрицательный, ЛП – ложноположительный, ЛО – ложноотрицательный.

Note. TP – true positive, TN – true negative, FP – false positive, FN – false negative.

Определим суммарные характеристики точности обнаружения в зависимости от *Dist_thresh* $\in$ [0,1] по всем изображениям с экспертной разметкой. Для каждого тэга возможны 4 исхода:

1) область обнаружена экспертом и детектирована алгоритмом – *TP* (истинно положительная ошибка);

2) область не обнаружена экспертом и не детектирована алгоритмом – *TN* (истинно отрицательная ошибка);

3) область обнаружена экспертом и не детектирована алгоритмом – *FN* (ложноотрицательная ошибка);

4) область не обнаружена экспертом и детектирована алгоритмом – *FP* (ложноположительная ошибка).

Правильное детектирование приватной области определяется степенью пересечения областей интереса. Если значение коэффициента Жаккара J(A,B) выше пороговой величины 0,7, то принимается факт правильного детектирования этой области:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

где A – область экспертной разметки, B – область, детектированная алгоритмом.

По всем экспертным разметкам для каждого порога *Dist_thresh* вычисляем чувствительность *TPR* и специфичность

$$TNR: TPR(Dist\_thresh) = TP/(TP + FN);$$
$$TNR(Dist\_thresh) = TN/(TN + FP);$$
$$FPR = 1 - TPR.$$

Из *TPR* и специфичности *FPR* строим *ROC*-кривую (receiver operating characteristic curve), представленную на рис. 5. Используя критерий Юдена, по *ROC*-кривой вычисляется оптимальный порог:

$$Dist\_optimal = argmax(TPR(Dist\_thresh) - FPR(Dist\_thresh)) \; \forall \; Dist\_thresh \in [0,1]).$$

По данному критерию мы получаем значение оптимального порога *Dist_optimal* = 0,28.

Оценка точности работы алгоритма

Для оптимального порога имеются 1 ложноположительный и 10 ложноотрицательных случаев работы алгоритма при 7917 объектах для поиска, содержащихся в 1131 изображении. Главной причиной этих случаев являются существенные ошибки первичного OCR. Матрица перепутывания представлена в таблице.

Итоговая точность детектирования составляет 99,86%. Метод легко адаптируется на другие списки служебных слов и параметров, содержащихся в изображении, что позволяет применять его на других классах медицинских изображений.

Обсуждение

В данной работе представлен метод распознавания текста, содержащего персональные данные на вспомогательных медицинских изображениях. Подготовлена база данных из 1131 медицинского изображения типа “Dose Report”, для которых эксперты разметили области с приватной информацией. На этой базе протестирован разработанный метод. Определен оптимальный порог принятия

решения о нахождении области личной информации. Алгоритм показал точность 99,86% на используемом наборе данных. Для аналогичных задач в литературе были показаны результаты точности распознавания 94–96% [3–7].

Дальнейшим развитием алгоритма является расширение типов и числа медицинских изображений, на которых его можно применять и тестировать.

Данный инструмент критически необходим при подготовке наборов медицинских данных для обучения и тестирования диагностических алгоритмов на основе искусственного интеллекта. Для этих целей наборы данных в анонимизированном виде передаются разработчикам. В последнее время также создаются облачные хранилища медицинских наборов данных для совместного пользования [15, 16], при этом крайне важно соблюсти требования по защите персональных данных [17]. Рекомендуемые способы анонимизации в основном затрагивают работу с DICOM-тэгами, но на практике мы сталкиваемся с исследованиями, содержащими вшитую персональную информацию не только в тегах, но и на самом медицинском изображении. Именно на решение задач автоматизации контроля присутствия персональных данных в исследовании и направлено наше исследование. В настоящее время отсутствуют готовые системные решения для данной задачи. Использование коммерческих OCR-решений существенно затруднено, а в отдельных случаях невозможно из-за низкого исходного разрешения изображения. Решение, разработанное в рамках данной работы, может быть рекомендовано к практическому применению для контроля качества подготовленных наборов данных.

Заключение

Разработанный в рамках настоящего исследования инструмент позволяет выявлять персональные данные на цифровых медицинских изображениях с точностью 99,86%, что указывает на возможность его практического применения при подготовке наборов данных. Исследования, на которых обнаружены области с приватными данными, не допускаются к включению в открытые датасеты без дополнительной обработки.

Разработка независимого метода первичного распознавания символов на изображениях является важным направлением, которое позволит поднять точность детектирования.

Финансирование

Публикация подготовлена при поддержке гранта Российского научного фонда № 22-25-20231, https://rscf.ru/project/22-25-20231/.

Участие авторов

Новик В.П. – проведение исследования, анализ и интерпретация полученных данных, написание текста.

Кульберг Н.С. – концепция и дизайн исследования, writing text.

Арзамасов К.М. – концепция и дизайн исследования.

Четвериков С.Ф. – проведение исследования.

Хоружая А.Н. – обзор публикаций по теме статьи, подготовка и редактирование текста.

Козлов Д.В. – статистическая обработка данных.

Кремнева Е.И. – участие в научном дизайне, утверждение окончательного варианта статьи.

Authors' participation

Novik V.P. – conducting research, collection and analysis of data, writing text.

Kulberg N.S. – concept and design of the study, writing text.

Arzamasov K.M. – concept and design of the study.

Chetverikov S.F. – conducting research.

Khoruzhaya A.N. – review of publications, text preparation and editing.

Kozlov D.V. – statistical analysis.

Kremneva E.I. – participation in scientific design, approval of the final version of the article.

Список литературы [References]

1. dicomstandard.org [Internet]. Dicom standard: Current Edition [cited 2022 Aug 27]. Available from: https://www.dicomstandard.org/current.
2. Aryanto K.Y.E., Oudkerk M., van Ooijen P.M.A. Free dicom de-identification tools in clinical research: functioning and safety of patient privacy. *Eur. Radiol.* 2015; 25 (12): 3685–3695. http://doi.org/10.1007/s00330-015-3794-0
3. Daye D., Wiggins W.F., Lungren M.P. et al. Implementation of Clinical Artificial Intelligence in Radiology: Who Decides and How? *Special Rep. Radiol.* 2022; 305 (1): E62. http://doi.org/10.1148/radiol.229021
4. dclunie.com [Internet]. David Clunie's Medical Image Format Site: Dicomcleaner [cited 2022 Aug 23]. Available from: http://www.dclunie.com.
5. Cook T.S., Zimmerman S.L., Steingall S.R. et al. Radiance: An automated, enterprise-wide solution for archiving and reporting ct radiation dose estimates. *Radiographics*. 2011; 31 (7): 1833–1846. http://doi.org/10.1148/rg.317115048
6. Vcelak P., Kryl M., Kratochvil M., Kleckova J. Identification and classification of dicom files with burned-in text content. *Int. J. Med. Inform*. 2019; 126: 128–137. http://doi.org/10.1016/j.ijmedinf.2019.02.011.
7. Monteiro E., Costa C., Oliveira J.L. A de-identification pipeline for ultrasound medical images in dicom format. *J. Med. Syst*. 2017; 41 (5): 89. http://doi.org/10.1007/s10916-017-0736-1.
8. Kin G., Tsui W., Chan T. Automatic selective removal of embedded patient information from image content of dicom files. *Am. J. Roentgenol*. 2012; 198 (4): 769–772. http://doi.org/10.2214/AJR.10.6352
9. Smith R. An overview of the Tesseract OCR engine. Proc. in Int. Conference on Document Analysis and Recognition (ICDAR). 2007; 629–633. http://doi.org/10.1109/ICDAR.2007.56

10. Левенштейн В. Двоичные коды с исправлением выпадений, вставок и замещений символов. *Доклады АН СССР.* 1965; 163: 845–848.
Levenshteyn V. Binary codes with correction of dropouts, inserts and substitutions of characters. *Doklady USSR Academy of Sciences.* 1965; 163: 845–848. (In Russian)
11. Schulz K., Mihov S. Fast string correction with levenshtein automata. *IJDAR*. 2002; 5: 67–85. http://doi.org/10.1007/s10032-002-0082-8
12. github.com [Internet]. Center of Diagnostics and Telemedicine. Find Anomalies in Tomography. Medical images markup system [cited 2022 Aug 3]. Available from: https://github.com/Center-of-Diagnostics-and-Telemedicine/FAnTom.
13. Mason D. SU-E-T-33: Pydicom: An Open Source DICOM Library. *Medical Physics.* 2011; 38 (6, Part 10): 3493–3493. http://doi.org/10.1118/1.3611983
14. Bradski G. The OpenCV Library. Dr Dobb's Journal of Software Tools. 2000.
15. Павлов Н.А., Андрейченко А.Е., Владзимирский А.В., Ревазян А.А., Кирпичев Ю.С., Морозов С.П. Эталонные медицинские датасеты (MosMedData) для независимой внешней оценки алгоритмов на основе искусственного интеллекта в диагностике. *Digital Diagnostics.* 2021; 2 (1): 49–66. http://doi.org/10.17816/DD60635
Pavlov N.A., Andreychenko A.E., Vladzymyrskyy A.V. et al. Reference medical datasets (MosMedData) for independent external evaluation of algorithms based on artificial intelligence in diagnostics. *Digital Diagnostics.* 2021; 2 (1): 49–66. http://doi.org/10.17816/DD60635 (In Russian)
16. Morozov S.P., Gombolevskiy V.A., Elizarov A.B. et al. A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT scans. *Comput. Methods Programs Biomed.* 2021; 206: 106–111. http://doi.org/10.1016/j.cmpb.2021.106111.
17. О персональных данных: [федер. закон: принят Гос. Думой 8 июля. 2006 г.: по состоянию на 2 июля 2021 г.]. On personal data: [federal law: adopted by the State. Duma on July 8. 2006: Accessed 2 July 2021 (In Russian)

Для корреспонденции*: Хоружая Анна Николаевна – тел.: +7-977-423-32-78. E-mail: khoruanna69@yandex.ru

Новик Владимир Петрович – научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ города Москвы "Научно-практический клинический центр диагностики и телемедицинских технологий ДЗ города Москвы", Москва. https://orcid.org/0000-0001-9481-1637

Кульберг Николай Сергеевич – канд. физ.-мат. наук, старший научный сотрудник Института кибернетики и образовательной информатики Федерального исследовательского центра "Информатика и управление" РАН, Москва. https://orcid.org/ 0000-0001-7046-7157

Арзамасов Кирилл Михайлович – канд. мед. наук, руководитель отдела медицинской информатики, радиомики и радиогеномики ГБУЗ города Москвы "Научно-практический клинический центр диагностики и телемедицинских технологий ДЗ города Москвы", Москва. https://orcid.org/0000-0001-7786-0349

Четвериков Сергей Федорович – начальник сектора разработки систем внедрения медицинских интеллектуальных технологий отдела медицинской информатики, радиомики и радиогеномики ГБУЗ города Москвы "Научно-практический клинический центр диагностики и телемедицинских технологий ДЗ города Москвы", Москва. https://orcid.org/0000-0002-3097-8881

Хоружая Анна Николаевна – младший научный сотрудник отдела инновационных технологий ГБУЗ города Москвы "Научно-практический клинический центр диагностики и телемедицинских технологий ДЗ города Москвы", Москва. https://orcid.org/0000-0003-4857-5404

Козлов Дмитрий Владимирович – младший научный сотрудник отдела медицинской информатики, радиомики и радиогеномики ГБУЗ города Москвы "Научно-практический клинический центр диагностики и телемедицинских технологий ДЗ города Москвы", Москва. https://orcid.org/0000-0002-4647-7301

Кремнева Елена Игоревна – канд. мед. наук, ведущий научный сотрудник отдела инновационных технологий ГБУЗ города Москвы "Научно-практический клинический центр диагностики и телемедицинских технологий ДЗ города Москвы"; старший научный сотрудник ФГБНУ "Научный центр неврологии", Москва. https://orcid.org/0000-0001-9396-6063

Contact*: Anna N. Khoruzhaya –phone: +7-977-423-32-78. E-mail: khoruanna69@yandex.ru

Vladimir P. Novik – Researcher of the Department of Medical Informatics, Radiomics and Radiogenomics, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of Moscow Health Care Department, Moscow. https://orcid.org/0000-0001-9481-1637

Nicholas S. Kulberg – Cand. of Sci. (Phys.-Math.), Senior Researcher of the Institute of Cybernetics and Educational Informatics, Federal Research Center Computer Science and Control of the Russian Academy of Sciences, Moscow. https://orcid.org/ 0000-0001-7046-7157

Kirill M. Arzamasov – Cand. of Sci. (Med.), Head of the Department of Medical Informatics, Radiomics and Radiogenomics, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of Moscow Health Care Department, Moscow. https://orcid.org/0000-0001-7786-0349

Sergey F. Chetverikov – Head of the Sector for the development of systems for the introduction of medical intelligent technologies of the Department of Medical Informatics, Radiomics and Radiogenomics, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of Moscow Health Care Department, Moscow. https://orcid.org/0000-0002-3097-8881

Anna N. Khoruzhaya – Junior Researcher of the Department of Innovative Technologies, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of Moscow Health Care Department, Moscow. https://orcid.org/0000-0003-4857-5404

Dmitriy V. Kozlov – Junior Researcher of the Department of Medical Informatics, Radiomics and Radiogenomics, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of Moscow Health Care Department, Moscow. https://orcid.org/0000-0002-4647-7301

Elena I. Kremneva – Cand. of Sci. (Med.), Leading Researcher of the Department of Innovative Technologies, Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of Moscow Health Care Department; Senior Researcher, Research center of neurology, Moscow. https://orcid.org/0000-0001-9396-6063