

Г. У. Бектемысова[✉], Е. Ж. Ахмер, А. Сабденов, Г.С. Бакирова
International University of Information Technology, Алматы, Казахстан
E-mail: g.bakirova@iitu.edu.kz

РАЗРАБОТКА МОДЕЛИ ДЛЯ КЛАССИФИКАЦИИ ДОКУМЕНТА (НА ПРИМЕРЕ ПАСПОРТОВ)

Аннотация. В данной статье рассматриваются подходы к предварительной обработке данных, классификации документов, коррекции ориентации и детекции текстовых полей с использованием модели YOLO для извлечения метаданных из документов, удостоверяющих личность. Процесс начинается с предварительной обработки изображений [1], включая нормализацию и масштабирование для унификации входных данных, соответствующих требованиям моделей. Затем изображение передается в модель классификации документов, которая определяет, соответствует ли оно критериям искомого документа, выполняя функцию первого фильтра, предотвращающего обработку нежелательных изображений. Если изображение успешно проходит классификацию, модель ориентации корректирует его положение, обеспечивая правильную ориентацию текста для дальнейшей обработки. Модель YOLO используется для определения и локализации текстовых полей на изображении, включая заголовки, абзацы и другие важные сегменты, которые требуют распознавания.

Ключевые слова. Классификация, сегментация, ориентация, документ, обработка данных, модель.

Введение.

Современные технологии обработки данных играют ключевую роль в автоматизации задач, связанных с идентификацией личности и проверкой документов. В условиях стремительного роста объема информации, требующей быстрой и точной обработки, особое внимание уделяется разработке систем, способных извлекать и анализировать данные из документов, удостоверяющих личность, таких как паспорта, водительские удостоверения и идентификационные карты. Эти системы востребованы в различных сферах, включая банковскую отрасль, транспорт, государственные услуги и системы безопасности [2].

Одной из ключевых задач в этой области является извлечение метаданных — структурированных данных, таких как имя, дата рождения, номер документа и другие важные сведения, из изображений документов. Традиционно для решения этой задачи используются технологии оптического распознавания символов (OCR), позволяющие преобразовать изображения текста в машинно-читаемый формат. Однако сложность современных документов, наличие нестандартных шрифтов, графических элементов и защитных функций значительно усложняют извлечение данных, что делает классические методы OCR недостаточно эффективными [3].

В последние годы значительное внимание уделяется внедрению методов глубокого обучения, которые продемонстрировали высокие результаты в задачах обработки изображений и естественного языка. Глубокие нейронные сети, такие как сверточные (CNN) и рекуррентные нейронные сети (RNN), а также модели на основе трансформеров, способны не только улучшить точность распознавания текста, но и извлекать структурированную информацию из документов даже в условиях сложных визуальных искажений и нестандартных форматов [4].

В нашем исследовании рассматриваются теоретические и методологические подходы к извлечению метаданных из документов, удостоверяющих личность, с использованием методов глубокого обучения [5] на основе оптического распознавания символов [6]. В статье будут обсуждены методы классификации, сегментации и ориентации полученных данных [7], а также ключевые аспекты, включая выбор архитектур нейронных сетей для задач OCR [8], методы предобработки изображений [9] и разработку моделей для классификации и интерпретации метаданных. Для обработки, классификации, сегментации, ориентации удостоверяющих личность нами было выбрано личный паспорт гражданина.

Материалы и методы.

В процессе обработки документов, удостоверяющих личность, важным этапом является эффективная сегментация и анализ изображения паспорта. Для достижения этой цели разработаны несколько ключевых моделей, каждая из которых выполняет свою уникальную задачу.

Модель сегментации отвечает за определение самого паспорта и его отделение от заднего фона. Этот этап критически важен для последующего анализа, так как точная сегментация позволяет минимизировать влияние фона на качество обработки.

Модель ориентации паспорта решает задачу определения угла поворота документа. Некоторые паспорта могут быть повернуты на 90 или 180 градусов, и правильная ориентация текста необходима для успешного распознавания информации.

Модель детекции полей осуществляет определение расположения всех полей на документе. Эта модель позволяет точно выявить области, содержащие важные данные, такие как имя, дата рождения и номер документа, что является основой для дальнейшего извлечения метаданных.

Совокупность этих этапов формирует надежный процесс обработки изображений паспортов, обеспечивая высокую точность извлечения информации и повышая эффективность систем автоматизации.

Модель классификации типов документов.

Сводное описание выборок.

Объем данных в обучающей, валидационной и тестовой выборках играет ключевую роль в разработке и оценке моделей машинного обучения. Он может существенно влиять на способность модели обобщать информацию и делать прогнозы для новых данных. В этом контексте количество данных в выборках можно считать достаточным для обучения и тестирования моделей (Таблица 1), если соблюдаются следующие условия.

Таблица 1- Сводное описание выборок

Характеристика	Обучающая выборка	Валидационная выборка	Тестовая выборка
Кол-во документов паспорта РФ	223	60	81
Кол-во других документов	282	67	54

Равномерное распределение: Хорошее представление различных классов или категорий в данных может способствовать более эффективному обучению и оценке моделей.

Сбалансированные выборки: Важно убедиться, что количество данных для каждого класса или категории в выборках сбалансировано. Несбалансированные выборки могут привести к искажению результатов и снижению способности модели к обобщению.

В нашем конкретном случае объем данных в обучающей, валидационной и тестовой выборках можно считать достаточным. У нас было разнообразные выборки с умеренным количеством документов каждого типа, и данные выглядят сбалансированными. Это создает условия для обучения и тестирования моделей с высокой вероятностью получения надежных результатов.

Задачей модели являлось – классифицировать тип документа, а именно паспорт. Данные, использованные для обучения модели, были обработаны в строгом соответствии с нормами и законодательством, регулирующими защиту персональных данных и обеспечение конфиденциальности.

Препроцессинг данных.

Процедуры предобработки (рисунок 2), которые были применены для каждого вида датасета (обучение, валидация, тестирование) описаны следующим образом.

```
input_size = (224, 224)

# Image transformation procedure
input_normalization = {'mean': mean.numpy(), 'std': std.numpy()}
train_transform = T.Compose([
    T.Resize(input_size),
    T.RandomVerticalFlip(),
    T.RandomPerspective(distortion_scale=0.6, p=0.7),
    T.RandomRotation(degrees=(0, 359)),
    T.ToTensor(),
    T.Normalize(**input_normalization)
])

val_transform = T.Compose([
    T.Resize(input_size),
    T.RandomVerticalFlip(p=0.7),
    T.RandomPerspective(distortion_scale=0.6, p=0.7),
    T.RandomRotation(degrees=(0, 360)),
    T.ToTensor(),
    T.Normalize(**input_normalization)])
```

Рисунок 2 - Препроцессинг данных (фрагмент Python Code)

Resize - изменение размера всех изображений до одного и того же размера, что является стандартной практикой для обучения нейронных сетей.

RandomVerticalFlip - случайное вертикальное отражение изображений, что может быть неидеальным для документов, так как оно может сделать текст нечитабельным.

RandomPerspective и RandomRotation - добавление случайных перспективных искажений и вращений может помочь модели быть более устойчивой к реальным условиям сканирования документов.

ToTensor и Normalize - преобразование изображений в тензоры и их нормализация, что является необходимым шагом перед подачей данных в нейронную сеть.

Для тестовой выборки использовалась только процедура ресайзинга.

Формирование выборок.

Для формирования выборок был использован словарь, созданный с помощью инструмента разметки Label Studio. В этом словаре имя файла выступает в качестве ключа, а координаты четырехугольника, охватывающего все пиксели, относящиеся к интересующему классу, являются значением. Таким образом, на вход датасету передаются путь к исходному изображению и путь к соответствующей метке класса ('pass', 'other').

Для построения модели использовалась библиотека pytorch и встроенные в нее пакеты.

Архитектура модели.

В качестве базовой архитектуры был использован предобученный UNet (Рисунок 3) с архитектурой MobileNet_v3, предварительно обученный на наборе изображений ImageNet. Подробная настройка архитектуры описана далее.

```
model = models.mobilenet_v3_large(pretrained=True)
model.classifier = nn.Sequential(
    nn.Linear(960, 1280),
    nn.Hardswish(),
    nn.Dropout(p=0.6, inplace=True),
    nn.Linear(1280, len(unique_labels))
)
```

Рисунок 3- Архитектура модели (фрагмент Python Code)

Фрагмент кода на изображении демонстрирует использование библиотеки PyTorch для модификации слоя классификатора предобученной модели MobileNet v3 [10]. В качестве функции потерь применялась встроенная torch.nn.BCEWithLogitsLoss(out, labels). Оптимизатором выбран torch.optim.Adam с начальной скоростью обучения (learning rate) 0.0001. Обучение модели проводилось в течение 20 эпох.

Обучение модели.

Обучение модели проводилось на GPU в течение 5 фолдов. Для мониторинга использовалась метрика accuracy. Для обеспечения стабильности модели была применена кросс-валидация с использованием метода KFold. Этот метод реализует процесс кросс-валидации для оценки производительности модели машинного обучения. Разделение данных осуществлялось с помощью KFold из библиотеки scikit-learn. Обучение и оценка проводились на каждом фолде, после чего результаты агрегировались для дальнейшего анализа.

Параметры:

- n_splits (int) - количество фолдов для разделения данных.
- shuffle (bool, по умолчанию True) - уремешивать ли данные перед разделением.
- random_state (int или None, по умолчанию None) - управляет случайностью разбиения данных.

Возвращаемым значением является среднее значение точности на валидации по всем фолдам. В итоге обучение прошло успешно, и модель продемонстрировала хорошие признаки обобщения.

Таблица 2 - Метрики качества классификации

	precision	recall	f1-score	support
other	1.00	0.99	0.99	81
pass	0.98	1.00	0.99	54
accuracy			0.99	135
macro avg	0.99	0.99	0.99	135
weighted avg	0.99	0.99	0.99	135

На изображении показана таблица с метриками качества классификации (Таблица 2), которая включает такие показатели, как точность (precision), полнота (recall), f1-оценка (f1-score) и поддержка (support).

1. Точность (precision) для категории "other" составляет 1.00, что означает, что все объекты, классифицированные моделью как "other", действительно принадлежат этому классу. Для категории "pass" точность равна 0.98, что указывает на то, что 98% объектов, отнесенных моделью к классу "pass", действительно относятся к этому классу.

2. Полнота (recall) для категории "other" составляет 0.99, что свидетельствует о том, что модель обнаружила 99% всех реальных объектов данного класса. Полнота для категории "pass" равна 1.00, что означает, что модель успешно идентифицировала все объекты этого класса.

3. F1-балл (f1-score), являющийся гармоническим средним между точностью и полнотой, одинаков для обеих категорий и составляет 0.99.

4. Поддержка (support) отражает количество реальных случаев в каждом классе, использованных для расчёта метрик: 81 для "other" и 54 для "pass".

В нижней части таблицы приведены средние значения по всем классам.

1. Accuracy (точность) отражает долю правильно классифицированных объектов и составляет 0.99.

2. Macro avg (макросреднее) представляет собой среднее арифметическое значений метрик по всем классам, равное 0.99 для f1-балла, 0.99 для точности и 0.99 для полноты.

3. Weighted avg (взвешенное среднее) учитывает количество объектов в каждом классе при расчете средних значений и также составляет 0.99 для f1-балла, точности и полноты.

Результаты.

В целом, высокие значения метрик свидетельствуют о том, что модель классификации показывает хорошие результаты для обоих классов. Итоговые метрики для тестовых выборок представлены на Рисунке 4.

На рисунке - 4 показана матрица ошибок (confusion matrix) для двух классов: "pass" и "other". Матрица ошибок - это способ визуализации производительности алгоритма классификации. В данной матрице:

1. По вертикальной оси (True labels) представлены истинные метки классов.

2. По горизонтальной оси (Predicted labels) представлены метки классов, предсказанные моделью.

В левом верхнем квадрате число 54 указывает на количество правильных предсказаний модели для класса "pass" (истинно-положительные результаты).

В правом нижнем квадрате число 80 обозначает количество правильных предсказаний для класса "other" (истинно-отрицательные результаты).

В левом нижнем квадрате число 1 указывает на количество случаев, когда модель неправильно предсказала класс "pass" для объектов истинного класса "other" (ложно-положительные результаты). Это также называют ошибками первого рода.

Правый верхний квадрат пуст, что означает, что не было случаев, когда модель ошибочно предсказала класс "other" для объектов истинного класса "pass" (ложно-отрицательные результаты или ошибки второго рода).

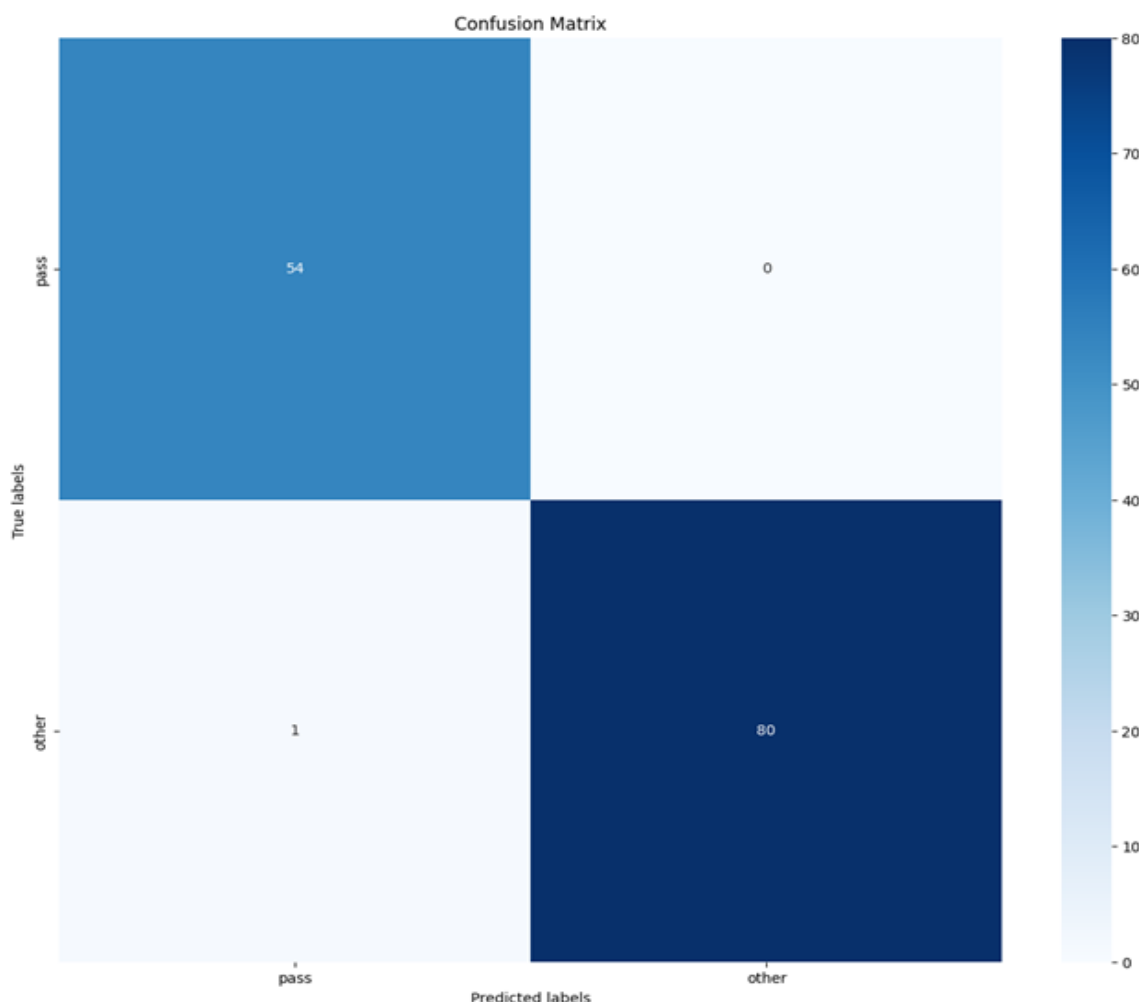


Рисунок 4 - Матрица ошибок

Матрица ошибок показывает, что модель демонстрирует хорошие результаты предсказания, особенно поскольку нет ложно-отрицательных предсказаний для класса "pass" и только четыре ложно-положительных предсказания для класса "other". Итоговая точность – 99,26%.

Обсуждение.

В процессе разработки и оценки модели классификации ключевую роль сыграло количество данных в обучающей, валидационной и тестовой выборках. Достаточный объем данных, их равномерное распределение и сбалансированность между классами обеспечили возможность успешного обучения модели [11]. Важно отметить, что соблюдение этих условий позволило избежать возможных искажений в результатах и обеспечило высокую точность модели. Несбалансированные выборки могли бы привести

к снижению качества обобщения модели и негативно сказаться на её способности делать точные прогнозы для новых данных.

Процесс обучения проходил на основе предобученной архитектуры UNet [12] с использованием MobileNet_v3 [6], что обеспечило высокую производительность при сравнительно малом количестве данных. Применение кросс-валидации с использованием метода KFold также способствовало стабильности модели и улучшению её обобщающей способности, что подтвердили результаты тестирования на всех фолдах. Качество работы модели было оценено с использованием метрик, таких как точность, полнота, f1-балл, которые показали высокие значения для обоих классов, что свидетельствует о правильной работе алгоритма классификации.

Матрица ошибок показала, что модель демонстрирует практически безошибочные результаты, с минимальными ложными срабатываниями. Только одно ложноположительное предсказание было зафиксировано для класса "pass", что указывает на высокий уровень достоверности модели. Отсутствие ложноотрицательных результатов для класса "pass" и высокая точность в предсказаниях для класса "other" говорят о том, что модель успешно справляется с задачей классификации документов.

Заключение.

Процесс обучения модели классификации документов продемонстрировал успешные результаты благодаря сбалансированному подходу к выборкам, продвинутой архитектуре модели и использованию методов кросс-валидации. Высокие значения метрик точности, полноты и f1-баллов, а также минимальное количество ошибок подтверждают, что модель может быть эффективно использована для классификации документов, таких как паспорта.

Использование предобученной модели MobileNet_v3 позволило существенно ускорить процесс обучения и достичь высоких результатов при сравнительно небольшом объеме данных. Результаты работы модели показывают, что она готова для практического применения и может быть успешно интегрирована в систему для автоматической обработки документов.

ЛИТЕРАТУРА

[1] J. Xu, F. Pan, X. Han, L. Wang, Y. Wang and W. Li, EdgeTrim-YOLO: Improved Trim YOLO Framework Tailored for Deployment on Edge Devices, 2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI), Xi'an, China, 2024, pp. 113-118, doi: 10.1109/CCAI61966.2024.10602964.

[2] K. V. Horadi, Document Image Analysis in Compressed Domain-Limitations, Applications & Challenges, 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 987-992, doi: 10.1109/ICECA49313.2020.9297593.

[3] V. N. Sai Rakesh Kamisetty, B. Sohan Chidvilas, S. Revathy, P. Jeyanthi, V. M. Anu and L. Mary Gladence, "Digitization of Data from Invoice using OCR, 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 1-10, doi: 10.1109/ICCMC53470.2022.9754117.

[4] C. Junliang, "CNN or RNN: Review and Experimental Comparison on Image Classification," 2022 IEEE 8th International Conference on Computer and Communications (ICCC), Chengdu, China, 2022, pp. 1939-1944, doi: 10.1109/ICCC56324.2022.10065984.

[5] A. R. F and L. Jacob, "Optical Character Recognition system with Projection Profile based segmentation and Deep Learning Techniques, 2022 4th International Conference on

Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2022, pp. 12-16, doi: 10.1109/ICAC3N56670.2022.10074151.

[6] P. Imsamer, V. Boonyaphon and S. Tiacharoen, "The Comparison of Deep Learning Driven Optical Character Recognition for Hard Disk Head Slider Serial Number," 2020 International Conference on Power, Energy and Innovations (ICPEI), Chiangmai, Thailand, 2020, pp. 217-220, doi: 10.1109/ICPEI49860.2020.9431431.

[7] S. K. Manocha and P. Tewari, "Comparative Study of Deep Learning Models for Devanagari OCR," 2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON), Pune, India, 2021, pp. 1-7, doi: 10.1109/SMARTGENCON51891.2021.9645924.

[8] C. Tensmeyer, D. Saunders and T. Martinez, "Convolutional Neural Networks for Font Classification," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 985-990, doi: 10.1109/ICDAR.2017.164.

[9] J. Chen, F. Li, Y. Fu, Q. Liu, J. Huang and K. Li, "A study of image segmentation algorithms combined with different image preprocessing methods for thyroid ultrasound images," 2017 IEEE International Conference on Imaging Systems and Techniques (IST), Beijing, China, 2017, pp. 1-5, doi: 10.1109/IST.2017.8261449.

[10] A. Saenong, Z. Zainuddin and M. Niswar, "Identification of Poultry Reproductive Behavior Using Faster R-CNN with MobileNet V3 Architecture in Traditional Cage Environment," 2023 International Seminar on Intelligent Technology and Its Applications (ISITIA), Surabaya, Indonesia, 2023, pp. 456-461, doi: 10.1109/ISITIA59021.2023.10221017.

[11] H. Yu, L. Gao, H. Yu and A. Zhang, "Vision Transformer based UNet with Multi-Head Attention for Medical Image Segmentation," 2024 36th Chinese Control and Decision Conference (CCDC), Xi'an, China, 2024, pp. 1737-1741, doi: 10.1109/CCDC62350.2024.10587821.

[12] X. Henghui, F. Yushen and H. Keke, "Research and Design of an IoT Face Recognition System Based on MobileNet-V3 and ArcFace," 2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 2023, pp. 635-639, doi: 10.1109/ICAICA58456.2023.10405625.

Гульнара Бектемысова, т.ғ.к., профессор, International University of Information Technology, Алматы, Қазақстан, g.bektemisova@iitu.edu.kz

Ерасыл Ахмер, докторант, International University of Information Technology, Алматы, Қазақстан, y.akhmer@iitu.edu.kz

Айдос Сабденов, докторант, International University of Information Technology, Алматы, Қазақстан, a.sabdenov@iitu.edu.kz

Гульназ Бакирова, докторант, International University of Information Technology, Алматы, Қазақстан, g.bakirova@iitu.edu.kz

ҚҰЖАТТАРДЫ ЖІКТЕУ ҮЛГІСІН ӘЗІРЛЕУ (ТӨЛҚҰЖАТ ҮЛГІСІН ПАЙДАЛАНУ)

Андатпа. Бұл мақалада жеке басын қуәландыратын құжаттардан метадеректерді алу үшін YOLO үлгісін пайдаланып деректерді алдын ала өңдеу, құжатты жіктеу, бағдарды түзету және мәтін өрісін анықтау тәсілдері талқыланады. Процесс кескінді алдын ала өңдеуден басталады [1], оның ішінде қалыпқа келтіру және үлгіге сәйкес кіріс деректерін біріктіру үшін масштабтау. Содан кейін кескін құжаттың жіктелу үлгісіне жіберіледі, ол оның ізделетін құжаттың критерийлеріне сәйкес келетінін анықтайды, қажетсіз кескіндердің өңделуіне жол бермеу үшін бірінші сүзгі ретінде әрекет етеді. Егер кескін

классификациядан сәтті өтсе, бағдар моделі одан әрі өңдеу үшін мәтіннің дұрыс бағдарлануын қамтамасыз ете отырып, оның орнын түзетеді. YOLO үлгісі суреттегі мәтін өрістерін, соның ішінде тақырыптарды, абзацтарды және тануды қажет ететін басқа маңызды сегменттерді анықтау және локализациялау үшін пайдаланылады.

Түйінді сөздер. жіктеу, сегменттеу, бағдарлау, құжаттама, мәліметтерді өңдеу, модель.

Gulnara Bektemyssova, candidate of technical sciences, professor, International University of Information Technology, Almaty, Kazakhstan, g.bektemisova@iitu.edu.kz

Yerasyl Akhmer, doctoral student, International University of Information Technology, Almaty, Kazakhstan, y.akhmer@iitu.edu.kz

Aidos Sabdenov, doctoral student, International University of Information Technology, Almaty, Kazakhstan, a.sabdenov@iitu.edu.kz

Gulnaz Bakirova, doctoral student, International University of Information Technology, Almaty, Kazakhstan, g.bakirova@iitu.edu.kz

DEVELOPMENT OF A MODEL FOR DOCUMENT CLASSIFICATION (USING THE EXAMPLE OF PASSPORTS)

Abstract. This paper discusses approaches for data preprocessing, document classification, orientation correction and text field detection using the YOLO model to extract metadata from identity documents. The process starts with image preprocessing [1], including normalization and scaling to unify model-compliant input data. The image is then passed to a document classification model, which determines whether it meets the criteria of the document being searched, acting as a first filter to prevent unwanted images from being processed. If the image successfully passes the classification, the orientation model corrects its position, ensuring the correct orientation of the text for further processing. The YOLO model is used to identify and localize text fields in an image including headings, paragraphs and other important segments that require recognition.

Keywords. Classification, segmentation, orientation, dokumet, data processing, model.

Редакцияға түсті / Поступила в редакцию / Received 26.09.2024
Жариялауға қабылданды / Принята к публикации / Accepted 07.02.2025